Applying discrete choice experiments to recent topics in health economics

DISSERTATION

zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften
(Dr. rer. pol.)
durch die Fakultät Wirtschaftswissenschaften der
Universität Duisburg-Essen
Campus Essen

vorgelegt von

Name: Björn Sossong

Ort: Rockenhausen

Essen (2014)

Tag der mündlichen Prüfung: 4. Juli 2014

Erstgutachter: Prof. Dr. Stefan Felder

Zweitgutachter: Prof. Dr. Hendrik Schmitz

Contents

List of figures

List of tables

1 Introduction

1.1 Aim and scope

The aim of this cumulative doctoral thesis is to empirically investigate four research questions related to the field of health economics:

1. Does saving an identified life differ from saving a statistical life in terms of utility?
2. Does the willingness to pay (WTP) for a reduction in mortality risk depend on the initial level of risk?
3. Do patients and non-patients associate different levels of utility to consequences of rheumatoid arthritis (RA)?
4. Are there substantial differences between the results of a discrete choice experiment (DCE) and a best-worst scaling (BWS) task?

The first two topics are concerned with two specific aspects related to the economic evaluation of a life, namely its degree of identifiability and the (initial) level of mortality risk it is exposed to. As both questions are analyzed in the same framework the results are directly comparable to each other.[1] The third topic takes another perspective and explores if being affected from an illness has an effect on the perceived importance of its relevant consequences. While this approach is not directly related to the valuation of a life, it addresses potential discrepancies in preference structures that may arise due to different states of affection.

All mentioned research questions have in common that the existing literature is inconclusive with respect to providing a definite answer to them. In order to shed more light in these contexts, this thesis reports the first applications of the comparatively novel DCE method to each research question. Topic 4 is motivated by this thesis' strong methodological focus on DCEs. Although it has often been acknowledged in the literature that DCEs are a theoretically well founded and popular preference elicitation method in health economics (see e.g. de Bekker-Grob *et al.*, 2012; Louviere *et al.*, 2010), it has been argued that its inherent weaknesses can be mitigated by the BWS task, a recently developed method that is in essence similar to DCEs (Flynn, 2010; Flynn *et al.*, 2007). Thus, it seemed worthwhile to include the more methodological oriented investigation of topic 4

This thesis was written with LaTeX; particularly, a modified version of a template created by Torsten Richter (`http://tortools.de`) was used.

[1] See section 6 for a comparison of the individual studies' results.

in this dissertation. The studies related to the four research questions are organized in chapters 2 to 5. Due to the emphasis on DCEs in this thesis, a brief historical perspective of their emergence is provided in the next section. In section 1.3 the structure of this dissertation is outlined in more detail.

1.2 The emergence of DCEs

Assumptions about the choice behavior of agents are an integral part of economic theory and thus have a long tradition. In classical economics, Mill described an individual "[...] who desires to possess wealth, and who is capable of judging the comparative efficacy of means for obtaining that end" (Mill, 1836, p. 321). Critics of Mill's work used – and eventually coined – the term 'economic man' to refer to this abstract concept of human behavior (Persky, 1995). In modern microeconomics, the view on the 'economic man', or, more generally, on individual decision making behavior has evolved and is formally characterized by a set of preference relations or choice rules (see e.g. Mas-Colell *et al.*, 1995, Chapter 1). Despite its explanatory power, the microeconomic framework relies on an axiomatic characterization of individual choice behavior and many results of this framework are based on contested propositions, e.g. the transitivity of preferences (Tversky, 1969). Thus, it is not very surprising that the need for an empirical investigation of individual preference patterns was emphasized as early as in the 1950s (May, 1954).

However, one major challenge related to the ability of analyzing preference patterns prevented immediate progress in this regard: The established econometric methods at that time were aimed at explaining aggregate and continuous variables and were not appropriate to analyze discrete choice behavior. Yet, many decision scenarios on the individual level involve discrete choices from a finite set of alternatives (Train, 1986, Chapter 1). For instance, a consumer decides to either rent or purchase a home. It was McFadden's random utility theory (RUT) that overcame this challenge and enabled researchers to empirically investigate discrete choice behavior on a solid theoretical foundation (McFadden, 1974). In contrast to earlier probabilistic discrete choice models formulated by Thurstone (1927) and Luce (1959), RUT treats individual choice as deterministic and can be used to derive the multinomial logit (MNL) model in order to conveniently estimate probabilities for discrete dependent variables.[2] Note that the version of the MNL model that only includes alternative specific regressors is sometimes referred to as the conditional logit (CLOGIT) model, its original name coined by McFadden (1974). Yet, as the MNL model had already been proposed before McFadden's formulation of RUT (Theil, 1969) and the difference between estimating a CLOGIT and a MNL model is nowadays mainly a practical issue related to the way the data is structured, I follow the majority of researchers in the choice analysis literature and refer to McFadden's version as MNL.

While McFadden's groundbreaking work provided a framework for the empirical analysis of discrete choices, the idea of Louviere (1973) and Davidson (1973) to use combinations

[2]See section 2.2 for a more detailed description of RUT and its relationship with the MNL model.

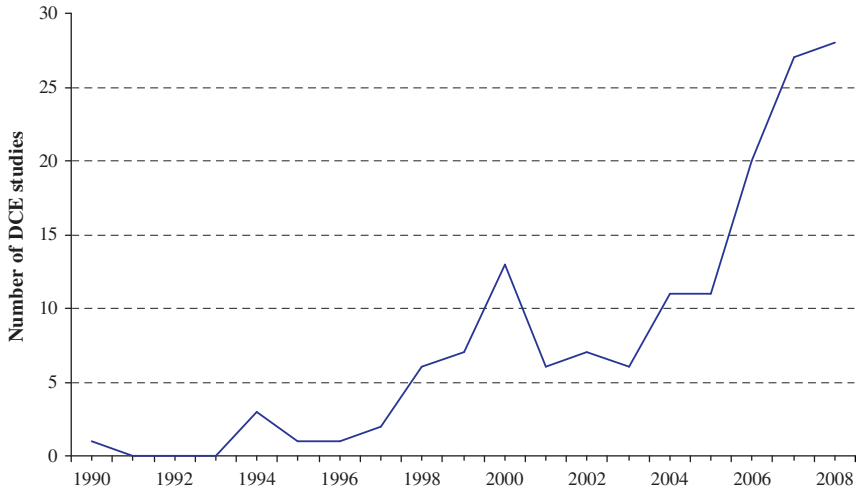

Figure 1.1: DCE studies by year of publication. Taken from de Bekker-Grob *et al.* (2012).

of levels of attributes to create hypothetical controlled choice situations allowed to transfer this framework to the realm of stated preference and investigate choice behavior in the absence of observable markets.[3] Although the essential theories had already been developed in the 1970s, the first reported DCEs with underlying experimental designs were performed by Louviere & Hensher (1982) and Louviere & Woodworth (1983). Due to the discrete nature of mode choice analysis and as the research focus of Louviere and Hensher was (and still is) transport economics, it is no surprise that DCE studies started to emerge in this branch of economics first.[4] Researchers in the field of environmental economics were also early adopters of the DCE method (Adamowicz *et al.*, 1994; Hoyos, 2010). DCEs were introduced in health economics in the early 1990s and the number of publications featuring DCEs has since increased by a remarkable margin (Ryan & Gerard, 2003). Figure 1.1 illustrates this development by depicting the number of DCE studies in health economic contexts by year of publication. It was taken from a recent systematic literature review by de Bekker-Grob *et al.* (2012).

Since their inception DCEs have undergone a considerable development with regard to their methodological aspects. In particular, much attention has been paid to two characteristics. One of these is the underlying experimental design, i.e. the result of the procedure that translates the identified relevant attributes and levels into a number of choice sets. In an idealized world, in which respondents had enough time and possessed the cognitive skills to answer all possible attribute-level combinations, experimental designs would be superfluous. As this is not the case and respondents can rarely be burdened with answering so-called full factorial designs, fractional factorial designs need to be created.[5]

[3] It is noteworthy in this context that Lancaster (1966) had already discussed a model for consumer demand in which the consumers possess preferences about the characteristics of goods and not preferences about the goods per se.

[4] See e.g. Hensher (1994) for a more detailed historical perspective of the emergence of DCEs in transport economics.

[5] For instance, the DCE reported in chapter 2 includes only four attributes with three levels each but its full factorial design contains 6,561 choice sets.

There are many different approaches to create an experimental design that only contains a small subset of the full design. However, the most prominent kinds are either based on the criterion of orthogonality, which requires that the (coded) attribute levels within an alternative are not correlated with each other across choice sets (see e.g. Hensher *et al.*, 2007), or the criterion of maximal statistical efficiency, where the d-efficiency measure is used predominantly in the DCE related literature (de Bekker-Grob *et al.*, 2012).[6]

Estimation procedures are the second important aspect of DCEs that has evolved over the years. The majority of early studies employed McFadden's MNL model with alternative specific regressors to estimate the utility differences between the levels in each attribute domain (de Bekker-Grob *et al.*, 2012). However, the MNL suffers from its strong assumptions of the irrelevance of independent alternatives (IIA) and independent and identically distributed (i.i.d.) extreme value type I error terms (see e.g Keane & Wasi, 2013). To mitigate these issues, so-called latent class models have been used that assume a discrete distribution for the error terms; however, it was argued that these models underestimate the degree of heterogeneity of choice data (Allenby & Rossi, 1998; Elrod & Keane, 1995). Therefore, other, more sophisticated, models have been proposed, such as the mixed logit (MIXL) model by Revelt & Train (1998) and, more recently, the scaled multinomial logit (SMNL) and generalized multinomial logit (GMNL) models by Fiebig *et al.* (2010).[7] In contrast to the MNL, these models are able to account for scale or taste heterogeneity across respondents. Although DCEs have become very popular and more sophisticated during the last two decades, there is anecdotal evidence that the interest in standard applications of DCEs has somewhat diminished, at least in the health economic research community. For instance, in the author guidelines of the journal Health Economics it says: "As a rule, the Journal does not include routine applications of cost-effectiveness analysis, discrete choice experiments and costing analyses" (Health Economics, 2013). Against this background, I will briefly summarize the content of the four studies that are contained in this cumulative dissertation thesis and specifically emphasize the 'non-routine' elements.

1.3 The four studies

Overview

The four studies that comprise this thesis are organized in the following chapters 2 to 5. Among other things, all studies have in common that they feature a DCE. However, in terms of content, different topics in health economics are investigated. In particular, the studies presented in chapters 2 and 3 build on the same data set and are concerned with topics related to the value of a statistical life (VSL): the identifiable victim effect and the dead-anyway effect. The study in chapter 4 investigates differences between patients and non-patients with regard to the evaluation of consequences of rheumatoid arthritis. These

[6] See Johnson *et al.* (2013) for a comprehensive discussion of prominent experimental design approaches.

[7] See chapter 2 for an application of the MIXL model and chapter 5 for applications of all mentioned models.

three studies share the novelty that a DCE is applied to a research question that has not been investigated with this method before. In the study presented in chapter 5, the data set used in chapter 4 is supplemented with a BWS survey, so that the results of the two rather similar methods can be compared with respect to substantial differences.

I am the sole author of 'Rescuing Schelling's girl' (chapter 2). 'The dead-anyway effect from a societal perspective' (chapter 3) is a joint work with Stefan Felder. 'Evaluating the consequences of rheumatoid arthritis' (chapter 4) originates from a collaboration with Stefan Felder, Malte Wolff and Klaus Krüger. Finally, 'A comparison of discrete choice and best-worst scaling' (chapter 5) is the result of a joint effort with Stefan Felder and Malte Wolff. The online questionnaires in all studies were programmed by Matthias Sieke.

Chapter 2: Rescuing Schelling's girl

It is widely believed that in modern societies identified lives are valued more highly than statistical lives. However, only a limited number of empirical studies finds evidence for the existence of this so-called identifiable victim effect. In this chapter, the results of a labeled DCE are reported, which was designed to overcome some of the limitations of the methods that have been applied to date to investigate preferences in the context of identified and statistical lives. We find that the respondents prefer reducing the mortality risk of an identified life over the one of a statistical life on an aggregate level. When expressed in monetary terms, the respondents attach an additional value of approximately USD 3.8 million to saving an identified life over an otherwise comparable statistical life. However, an analysis of subgroups suggests that the older part of the respondents in the sample drives this effect.

The novelty in this study's approach lies in the identification strategy of the identifiable victim effect: By using effects coded variables the alternative specific constant (ASC) in the estimated models can be used to test the significance of the identifiable victim effect and even calculate the difference in value between an identified and a statistical life.

Chapter 3: The dead-anyway effect from a societal perspective

Drawing on the same data set used in chapter 2, we use more sophisticated MIXL models to investigate the relationship between willingness to pay (WTP) for a given level of risk reduction and the initial mortality risk of the beneficiary when society spends the resources. Pratt & Zeckhauser (1996) assert that this relationship should be positive on the individual level and termed it the dead-anyway effect (DAE). We find evidence in favor of this assertion on the societal level. Furthermore, our results suggest that the WTP for a fixed risk reduction does not seem to increase incrementally with initial risk.

This is the first study to our knowledge that uses a DCE to analyze if the DAE can be observed. In addition, the study departs from the original assertion of Pratt & Zeckhauser (1996) as the existence of the DAE is investigated on the societal instead of on the

individual level, which may be better suited to derive policy implications. The method also allows to calculate a theoretically founded value for the WTP for initial risk reductions.

Chapter 4: Evaluating the consequences of rheumatoid arthritis

Patients and non-patients tend to attach different utility values to the state of suffering from specific illnesses (Ubel *et al.*, 2000). This observation naturally leads to the question whose utility values should be used as the basis in cost-effectiveness analysis (CEA). Intuitively, one would presume that patients are better informed about the consequences of their illness and public authorities should therefore use the patients' utility values in CEA. Contrary to this presumption, it has been argued that society at large should determine which values are to be used and not the patients because, in the end, it is societal resources that are to be allocated. Against this background, we use data from a DCE that was completed by patients suffering from RA and non-patients to explore the discrepancies between the two groups' utility estimates for typical consequences of RA. Our results indicate that both groups attach remarkably similar part-worth utilities to the symptoms pain, fatigue and functional limitations. However, non-patients significantly undervalue the ability to work when compared to patients.

This study is novel in two aspects: First, the analysis of patients and non-patients within the DCE framework enables us to quantify the differences in the effects of the selected underlying characteristics of RA. Prior research has focused on the estimation of differences between utility values of whole health states. Second, we explicitly investigate whether being incapacitated for work is perceived to be of different importance by patients and non-patients.

Chapter 5: A comparison of discrete choice and best-worst scaling

In this study, we supplement the DCE described in the previous chapter by a BWS survey with the identical underlying experimental design to shed more light on potential differences between the two methods. In particular, we estimate models capable of accounting for taste and scale heterogeneity in order to investigate if there are method induced differences in this regard. We find that the DCE and BWS lead to considerably different coefficient estimates and that the BWS responses exhibit a larger degree of taste heterogeneity.

This is an early application of the GMNL model and one of few studies that compare DCE and BWS results. While this work is in its structure similar to the approach of Whitty *et al.* (2013), we take a more explicit look at the differences between the two methods with regard to taste and scale heterogeneity.

2 Rescuing Schelling's girl

2.1 Introduction

"There is a distinction between an individual life and a statistical life. Let a six-year-old girl with brown hair need thousands of dollars for an operation that will prolong her life until Christmas, and the post office will be swamped with nickels and dimes to save her. But let it be reported that without a sales tax the hospital facilities of Massachusetts will deteriorate and cause a barely perceptible increase in preventable deaths – not many will drop a tear or reach for their checkbooks." (Schelling, 1968)

In this much-cited passage Schelling asserts that people are willing to pay more for a reduction in risk to an individual or identified life than to a statistical life. Although Schelling's assertion appears to be widely accepted as true (Hammitt & Treich, 2007), most studies refer only to specific cases of identified lives at risk to provide evidence in favor of the described effect. In the contexts of these specific cases it is often remarked that the resources used to save the identified lives could have been used better to save – sometimes many more – statistical lives (Lamm, 2001; Moore, 1996; Richardson & McKie, 2003). For instance, Jenni & Loewenstein (1997) mention the case of "baby Jessica", in which the family of a baby trapped in a well received over USD 700,000 in donations although this amount of money could have been used to save more statistical lives.

Even though Schelling's assertion may be convincing, one should be cautious not to conclude solely from such exemplary evidence that people value identified lives more than statistical lives. From an economist's point of view, people are likely to have only incomplete information about the opportunity cost of providing help to a specific identified life at risk. If "baby Jessica's" donors had been given the information that their contribution could help many more statistical lives, then maybe they would have decided differently. This argument is supported by the findings of Small *et al.* (2007), who report that the discrepancy between giving toward identified and statistical lives seems to disappear if potential donors are informed that people typically react more strongly to identified than statistical lives at risk and are given an example for this behavioral tendency. Against this background, it seems worthwhile to investigate whether people still prefer saving an identified life to saving a statistical life if they are explicitly made aware of the opportunity cost that are associated with their decision. In choice analysis this could, for example, be accomplished by presenting the respondents with a choice task that requires them to

I am the sole author of 'Rescuing Schelling's girl'. See Sossong (2012) for a published earlier working paper version of this chapter.

choose between helping either an identified or a statistical life, so that the information about the forgone alternative is readily available.

In addition to Small *et al.* (2007), there are four studies, to the best knowledge of the author, that empirically analyze Schelling's claim, which was termed the identifiable victim effect in this strand of literature. Using a laboratory experiment, Small & Loewenstein (2003) find that merely the information that the recipient of a potential donation has already been selected increases the willingness to contribute (WTC). Based on the results of a survey, Kogut & Ritov (2005a) report that the willingness to donate to an identified sick child is higher if a picture is disclosed to the potential donor. These two studies feature methods which do not force the respondents to make an explicit choice between helping either an identified or a statistical life. Rather, the authors randomly assign either case to the respondents and compare the two groups' average WTC. However, Kogut & Ritov (2005b) show that it seems to make a large difference whether the WTC values are elicited separately from different populations or jointly by presenting a choice task to one population. They find that while the respondents have a significantly higher WTC to a single victim than to a group of victims when evaluated separately, this preference reverses when the respondents are requested to choose between either contributing to the single victim or to a group of victims. Jenni & Loewenstein (1997) also employ, inter alia, a method that forces respondents to make a choice. They present pairs of scenarios and ask the respondents to select the one in which they find it most important to eliminate the risk. Surprisingly, the authors find that out of four factors that may potentially cause the identifiable victim effect only the proportion of the number of people that can be saved to the perceived reference group at risk is positively correlated with the WTC. This result implies that providing help in a scenario in which 10 out of 100 people can be saved is more important than providing help in a scenario in which 10 out of 100,000 people can be saved.

Thus, although there is compelling intuitive evidence that the identifiable victim effect exists, empirical studies have been able to convincingly support its existence only if the respondents are not informed about their decisions' opportunity cost. Studies in which the respondents have information about opportunity cost, e.g. if trade-off techniques are employed to elicit WTCs, report either only weak empirical support (Jenni & Loewenstein, 1997), no support at all (Small *et al.*, 2007) or even a reversal in preferences (Kogut & Ritov, 2005b). This paper builds on these inconclusive empirical findings and reports the results of a study in which a labeled discrete choice experiment (DCE) was used to analyze whether German citizens differentiate between the value assigned to a reduction in the mortality risk of a statistical life and an identified life.

There are four major reasons why a DCE may add to gain a better understanding of the identifiable victim effect. First, in a DCE people can be forced to take decisions while providing them with information about their decision's opportunity cost. Second, DCEs have a theoretical foundation for the elicitation of preferences. Third, the respondents' willingness to pay (WTP) for saving an identified over a statistical life can be inferred from the estimated coefficients. Fourth, in the DCE framework one can control for all given personalizing information on the victims by including the relevant parameters in

the model specification. A DCE can thus overcome the problem of the hitherto applied methods that people may respond to the information that is given rather than to the identification per se. This has been noted as one major challenge in demonstrating the existence of the identifiable victim effect (Small & Loewenstein, 2003).

In the employed DCE, the distinct notions of a statistical and an identified life are operationalized by introducing two labeled alternatives: a preventive program that provides benefits to an unspecified individual and a curative program that provides benefits to a diseased individual. Although the exact wording of the labels is to some degree arbitrary, it is based on the findings of the identifiable victim effect literature so that the two labels resemble the two distinct notions of an identified and a statistical life as closely as possible. It is worth mentioning at this point that comparing the general public's preferences with regard to a preventive and a curative program is of high practical relevance. Pratt & Zeckhauser (1996) note in this context that "it is often alleged that our society devotes an imbalance of resources toward treatment after the fact as opposed to prevention [...]".

The remainder of this study is organized in four sections. In the following section 2.2 the theoretical underpinnings of DCEs are briefly summarized. Subsequently, the applied DCE is presented in section 2.3 The results of the estimation are shown in section 2.4. Finally, a concluding discussion is provided in section 2.5.

2.2 DCE framework

DCEs are a stated preference technique. In a DCE, the respondents are presented with a number of choice sets composed of at least two competing alternatives that vary along specified attributes. As the estimated complete ordering of relevant preferences is subtly elicited through a number of discrete choices, DCEs demand comparatively weak assumptions about human cognitive abilities (Louviere *et al.*, 2000). This advantage and others, including the ability to infer marginal rates of substitution across monetary and non-monetary attributes, may contribute to explaining the popularity of studies using DCEs in various fields of the economic literature, such as in health economics (de Bekker-Grob *et al.*, 2012), environmental economics (Hoyos, 2010) and transport economics (Hensher, 1994). DCEs have also been used in other contexts, e.g. to elicit preferences that allow the calculation of the value of a statistical life (VSL) (Tsuge *et al.*, 2005) and to investigate people's preferences for live theater (Grisolía & Willisa, 2011).

The theoretical foundation of DCEs draws upon modern microeconomic consumer theory, Lancaster's argument that it is the attributes of goods that determine the utility they provide (Lancaster, 1966) and random utility theory (RUT) (McFadden, 1974). RUT assumes that the overall utility of the ith alternative for the nth individual U_{in} is additively and independently composed of the observable source of utility V_{in} and the unobservable source ε_{in}. This can be written as

$$U_{in} = V_{in} + \varepsilon_{in}. \tag{2.1}$$

Furthermore, it is assumed that the observable component is a function of the vector of the alternative's attributes x_{in}, so that

$$V_{in} = V_{in}(x_{in}). \tag{2.2}$$

In most DCE applications, the observable component is simply specified as a weighted linear expression with $k = 1, ..., K$ attributes that enter the the utility function V_{in} in a customizable functional form f:

$$V_{in} = \beta_{0i} + \beta_{1i} f(x_{1in}) + \beta_{2i} f(x_{2in}) + \beta_{3i} f(x_{3in}) + ... + \beta_{Ki} f(x_{Kin}). \tag{2.3}$$

The parameter β_{0i} is called the alternative-specific constant and $\beta_{1i}, ..., \beta_{Ki}$ are the weights associated with attribute k of alternative i. Most studies treat attributes as linear and define $f(x) = x$. However, some studies also account for interactions between attributes or specify them in a logarithmic form or as a quadratic. To derive a choice model that is capable of estimating the aforementioned attribute weights, it is assumed that the individual will compare all alternatives within a choice set $j = 1, ..., i, ...J$ according to their overall utility U_{jn} and choose the one with the maximum utility. As the overall utility of an alternative includes an unobservable and therefore probabilistic component from the point of view of the analyst, the individual's choice behavior is explained in terms of probabilities. Specifically, it is defined that the probability that an individual n will choose alternative i is given by:

$$\begin{aligned} Prob_{in} &= Prob[(U_{in} \geq U_{jn})] \\ &= Prob[(V_{in} + \varepsilon_{in}) \geq (V_{jn} + \varepsilon_{jn})] \qquad \forall j \in j = 1, ..., J; i \neq j \\ &= Prob[(\varepsilon_{jn} - \varepsilon_{in}) \leq (V_{in} - V_{jn})]. \end{aligned} \tag{2.4}$$

For reasons of simplicity and desirable distributional properties, it is assumed in many studies that the unobserved component ε is independently and identically extreme value type I distributed. With this assumption the multinomial logit (MNL) model can be derived:

$$Prob_{in} = \frac{exp(V_{in})}{\sum_{j=1}^{J} exp(V_{jn})}. \tag{2.5}$$

The MNL model can be used to estimate the attribute weights in the observable source of utility V_{in}. Due to its simplicity and empirical value, the MNL model has become very popular in discrete choice analysis and is frequently used in studies that employ DCEs (Hensher *et al.*, 2007; Louviere *et al.*, 2000). This is the basic model that is used to estimate the parameters of the DCE presented in this study.

2.3 Methods

The aim is to investigate whether German citizens prefer reducing the mortality risk of an identified life over reducing the mortality risk of a statistical life, so that the predicted behavior of the identifiable victim effect can either be supported or contested. Therefore, a labeled DCE with a final sample of 210 German citizens was conducted using an online questionnaire.

2.3.1 Setting

A review of the identifiable victim effect literature reveals that three differences between a statistical and an identified life at risk seem to be correlated with people's perceived urgency of providing help.

First, statistical and identified lives in danger seem to differ with regard to the size of the reference group. Intuitively, an identified life constitutes its own reference group of size one whereas the reference group of a statistical life is usually larger (Small *et al.*, 2007). As noted in section 2.1, empirical investigations suggest that the proportion of the number of people that can be saved to the reference group at risk seems to be positively correlated with people's valuation of lifesaving interventions (Fetherstonhaugh *et al.*, 1997; Jenni & Loewenstein, 1997). Second, the status of determination of the victim distinguishes identified from statistical lives at risk. In contrast to statistical victims, the occurrence of the risk-producing event usually lies in the past for identified victims. Small & Loewenstein (2003) find that the sole information that a victim is determined – without providing any further personalizing information – increases caring. Third, the knowledge of personalizing information of victims correlates with people's WTC (Kogut & Ritov, 2005a). Specifically, the disclosure of the victim's picture is of significance in this context.

Along these three differences, two distinctly labeled alternatives were defined. They were labeled 'preventive safety measures' and 'therapy for a diseased person'. The label 'preventive safety measures' reflects the traits of an intervention aimed at providing benefits to an endangered statistical life; it conveys that a not yet affected and unspecified person with a reference group larger than one is expected to receive health gains if this alternative is chosen. Analogously, the label 'therapy for a diseased person' was chosen to reflect a measure that aims at providing health benefits to an individual life; the label suggests that a person, who has already been determined to suffer, with a reference group of size one is at risk. Table 2.1 summarizes the differences between the two labels along the three distinction criteria found in the literature.

These two alternatives were embedded in a hypothetical choice context. The respondents were asked to imagine a scenario in which their opinion is requested to resolve a tie vote on the allocation of funds between two alternatives in a health economics expert committee. They were informed that it is only possible to fund one of two alternatives in each choice set and that by choosing an alternative the mortality risk of only one person,

Table 2.1: Differences between identified and statistical life

Distinction criterion	Alternative 1 Preventive safety measures	Alternative 2 Therapy for a diseased person
1. Proportion of people that can be saved to the reference group at risk (Jenni & Loewenstein, 1997)	Unspecified, but perceivably smaller than one	One (only diseased person)
2. Person at risk already selected (Small & Loewenstein, 2003)	No	Yes
3. Personalizing information given (Kogut & Ritov, 2005a)	Only broad category of recipients given	Yes, age of diseased person given

be it a preventive safety measure or a therapy for a diseased person, will be reduced by 50 percentage points for the next five years. The mortality risk of one person in the forgone alternative was said to not change within the next five years. It was emphasized that in both alternatives the given mortality risk is related to only one individual.

Admittedly, there are numerous ways of formulating the labels to distinguish the notions of an identified and a statistical life, and one might conclude that by using the significant empirical findings of the identifiable victim effect literature to derive these labels the experiment becomes a self-fulfilling prophecy. However, the novelty and advantage of this approach is that the respondents' preference structure regarding helping identified and statistical lives is elicited in the presence of opportunity cost. In the context of the finding that preferences may reverse in certain settings where opportunity cost are known (Kogut & Ritov, 2005b), it is unclear a priori how the respondents will decide. So, this study design forces people to make a choice between helping either an identified or a statistical life, where the notions of these two abstract concepts are formulated in accordance with the results of prior research.

2.3.2 Attributes and levels

Four attributes with three levels each were included in the DCE both to provide a realistic choice context for the respondents and to analyze their effect on choice behavior. Both alternatives shared the attributes 'mortality risk of one person' (RSK) and 'cost for decreasing the mortality risk by 50 percentage points' (CST). The levels of RSK were chosen as 50%, 75% and 100% to ensure equidistance and to convey that not choosing an alternative poses a considerable threat to one life. The CST attribute included the

Table 2.2: Attributes and levels

Attribute	Alternative	Levels
Area of investment (INV)	Preventive safety measures	Elementary schools (ELEM) Road junctions (ROAD) Nursing homes (NURS)
Age of diseased person (AGE)	Therapy for a diseased person	10-year-old person (AGE10) 40-year-old person (AGE40) 70-year-old person (AGE70)
Mortality risk of one person (RSK)	Both alternatives	50% (RSK50) 75% (RSK75) 100% (RSK100)
Cost for decreasing the mortality risk by 50 percentage points (CST)	Both alternatives	EUR 1 mio. (CST1) EUR 2.5 mio. (CST2.5) EUR 4 mio. (CST4)

levels EUR 1, 2.5 and 4 million. These values were chosen to be around half the typical range that Viscusi & Aldy (2003) report in their VSL investigation because in this case the mortality risk is reduced by only 50 percentage points (Tsuge *et al.*, 2005).

Furthermore, two alternative specific attributes were added. For the 'preventive safety measures' alternative the 'area of investment' (INV) was provided to give the respondents a broad reference group for the life at risk. The attribute could take the levels 'elementary schools', 'road junctions' and 'nursing homes'. For the 'therapy for a diseased person' alternative the age of the diseased person (AGE) with the levels 10-, 40- and 70-year-old person was given. For simplicity, these two attributes were presented to the respondents in the same row in each binary choice set, where the row was labeled 'description of allocation of funds' (DOF). Thus, in each choice set DOF could either take the values of INV if the alternative was labeled 'preventive safety measures' or the values of AGE if the alternative was labeled 'therapy for a diseased person'. Table 2.2 provides an overview of all used attributes and levels and figure 2.1 presents a sample choice set.

2.3.3 Presentation of choice sets

An orthogonal design was used to reduce the resulting 6,561 ($3^{2\times4}$) possible choice sets to 27. The chosen design ensures the absence of correlation among all main effects and provides the required number of degrees of freedom to estimate the parameters of interest. In addition, selected two-way interaction effects are not confounded with each other and with the main effects in the design, and thus could be included in the subsequent model specifications. Three rationality test choice sets were added to these 27 to be able to

Scenario 25	Therapy for a diseased person	Preventive safety measures
Description of allocation of funds	70-year-old person	Safety measures for elementary schools
Mortality risk of one person	100%	75%
Cost for decreasing the mortality risk by 50 percentage points	EUR 2.5 million	EUR 1 million
I decide in favor of	☐	☐

Figure 2.1: Translated example of a choice set (original in German)

exclude irrational respondents from the analysis. In these test choice sets, the respondents are confronted with two nearly identical alternatives that only differ with regard to the CST attribute, i.e. even the labels of the two alternatives were the same. This rationality test is in essence a test for non-satiation (Lancsar & Louviere, 2006; Miguel *et al.*, 2005). Furthermore, the sequence in which the two labeled alternatives were shown to the respondents was randomized to avoid an order effect. This means that the 'preventive safety measures' alternative sometimes appeared on the left-hand side and sometimes on the right-hand side in the questionnaire. Table 2.3 shows all choice sets used in the study.

2.3.4 Sample and data collection

A questionnaire was developed in HTML and JAVA and published online. The questionnaire consisted of three parts: an introduction, the 30 choice sets and conventional socio-economic questions, including two questions on the difficulty of understanding the task at hand and of making a choice. In the introduction, the nature of the study was explained and the choice context as well as descriptions of the attributes were provided. At every step of the DCE part of the questionnaire, the respondents could access the information given on the choice context and the attributes. Before the survey was carried out, a pretest with 36 master students at the University of Duisburg-Essen was conducted using a paper-based version of the questionnaire to test and improve its comprehensibility.

The sampling and data collection process was managed by a market research company using their online panel. 599 of the panel's total 4,437 participants were selected to participate in this study using quota sampling with regard to age group, sex, federal state (Bundesland) of residency and level of education. To improve representativity of the final sample, this was done in two waves: 312 participants were asked to fill out the questionnaire in the first wave and 287 in the second wave. In total, 367 people completed the questionnaire, implying a response rate of 61%. The respondents were incentivized as in comparable studies by receiving an amount of points that can be used to buy products in the institute's online shop. After the respondents logged on to their specific accounts in

Table 2.3: Summary of all choice sets

Choice set	Preventive safety measures			Therapy for a diseased person		
1	ELEM	RSK75	CST2.5	AGE10	RSK75	CST4
2	Rationality test[a]					
3	ROAD	RSK50	CST1	AGE70	RSK50	CST4
4	NURS	RSK75	CST4	AGE70	RSK50	CST2.5
5	NURS	RSK75	CST1	AGE10	RSK100	CST4
6	NURS	RSK100	CST2.5	AGE10	RSK50	CST2.5
7	NURS	RSK100	CST1	AGE70	RSK75	CST1
8	ELEM	RSK100	CST1	AGE40	RSK75	CST4
9	ROAD	RSK50	CST4	AGE40	RSK75	CST2.5
10	ROAD	RSK75	CST1	AGE40	RSK100	CST1
11	NURS	RSK50	CST2.5	AGE70	RSK100	CST4
12	ELEM	RSK50	CST4	AGE70	RSK75	CST4
13	ELEM	RSK75	CST4	AGE40	RSK50	CST1
14	ELEM	RSK100	CST4	AGE10	RSK100	CST2.5
15	Rationality test[a]					
16	NURS	RSK50	CST1	AGE40	RSK50	CST2.5
17	NURS	RSK75	CST2.5	AGE40	RSK75	CST1
18	ROAD	RSK75	CST2.5	AGE70	RSK75	CST2.5
19	ELEM	RSK100	CST2.5	AGE70	RSK50	CST1
20	ROAD	RSK100	CST4	AGE70	RSK100	CST1
21	ELEM	RSK50	CST2.5	AGE40	RSK100	CST2.5
22	ROAD	RSK50	CST2.5	AGE10	RSK100	CST1
23	ROAD	RSK75	CST4	AGE10	RSK50	CST4
24	ROAD	RSK100	CST2.5	AGE40	RSK50	CST4
25	ELEM	RSK75	CST1	AGE70	RSK100	CST2.5
26	ELEM	RSK50	CST1	AGE10	RSK50	CST1
27	NURS	RSK100	CST4	AGE40	RSK100	CST4
28	NURS	RSK50	CST4	AGE10	RSK75	CST1
29	ROAD	RSK100	CST1	AGE10	RSK75	CST2.5
30	Rationality test[a]					

[a] Apart from CST the two alternatives are identical.

Table 2.4: Selected descriptive statistics of the sample

	N (sample)	% (sample)	% (Pop.)[a]	△%
Sex				
Female	99	47.1	51.0	3.9
Residency				
Schleswig-Holstein	6	2.9	3.5	0.6
Hamburg	5	2.4	2.2	-0.2
Lower Saxony	16	7.6	9.7	2.1
Bremen	2	1.0	0.8	-0.1
Northrhine-Westphalia	41	19.5	21.9	2.3
Hesse	14	6.7	7.4	0.7
Baden-Wurttemberg	36	17.1	13.1	-4.0
Rhineland Palatinate	11	5.2	4.9	-0.3
Bavaria	35	16.7	15.3	-1.4
Saarland	2	1.0	1.3	0.3
Berlin	10	4.8	4.2	-0.6
Brandenburg	4	1.9	3.1	1.2
Mecklenburg Western Pomerania	4	1.9	2.0	0.1
Saxony	15	7.1	5.1	-2.0
Saxony-Anhalt	2	1.0	2.9	2.0
Thuringia	7	3.3	2.8	-0.6
Age				
<18	0	0.0	16.7	16.7
18 to <40	69	32.9	27.0	-5.9
40 to <60	84	40.0	30.8	-9.2
≥60	57	27.1	25.6	-1.6
Education				
No school leaving certificate	0	0.0	3.9	3.9
Cert. of Secondary Education	53	25.2	39.3	14.0
Gen. Cert. of Sec. Education	54	25.7	21.1	-4.6
Qual. for university entrance	93	44.3	24.4	-19.8
Other	5	2.4	10.3	8.0
Not specified	5	2.4	0.4	-1.9

[a] All figures are for the year of 2008 and are taken from Federal Statistical Office (2010), apart from the information on education, which is taken from Federal Statistical Office (2009) due to availability.

the company's system, they were instructed to follow a link that led them to the online questionnaire. This procedure made it possible to track the identification number of every respondent who accessed the questionnaire. Of the 367 respondents 26 were excluded from further analysis because they encountered technical problems while answering the questionnaire and thus did not complete all choice sets. 40% of the remaining 341 respondents did not pass all three rationality test choice sets; i.e. they did not choose the lower cost – of the otherwise identical – alternatives in at least one of the three tests. These respondents were excluded from further analysis because it was assumed that they either randomly chose alternatives to quickly complete the questionnaire and receive their points, lost their concentration or are completely insensitive to cost in the presented choice task. It is noteworthy that including these 'irrational' respondents in the MNL models does not considerably change the sizes of the estimated parameters and has barely any effect on their significance levels. In addition, if one only analyzes the sample of irrational respondents, the results are very similar to those of the rational sample in parameter sizes and significance levels, so that nearly all the same conclusions could be drawn. The only major difference between the 'irrational' and the 'rational' sample is that the former appears to be rather cost insensitive, as all estimated CST parameters are not significantly different from zero. Thus, the major results of this study appear to be robust regardless of how 'non-rational' respondents are treated.

For the remaining 210 rational respondents, who constitute the sample for the results shown in the following section, table 2.4 provides descriptive statistics of the characteristics that were used as quotas in the sampling process. There is only a slightly higher proportion of males in the sample than in the German population in 2008 (3.9 percentage points). In addition, the sample is representative of the population concerning residency on the state level, the maximum difference in terms of absolute percentage points being 4.0. Unfortunately, the market research institute's online panel does not include people under the age of 18, which explains the discrepancies in the age group category. Also, the sample includes a considerably larger number of respondents with the highest level of school education than the population at large (19.8 percentage points difference), which is mainly due to the online nature of the survey.

2.3.5 Data analysis

Four MNL models are employed to analyze the data. In model 1, all $3-1$ levels of the three-level attributes enter the indirect utility functions linearly, additively and effects coded. The levels NURS, AGE70, RSK100 and CST4 of the respective attributes are chosen as base levels. This model includes only main effects because all two-way interaction effects that were deemed potentially important during the experimental design stage turned out to be insignificant in subsequent analysis. As the DCE includes two labeled alternatives in each binary choice set, the part-worth utilities of all attribute levels are estimated separately for each alternative.[1] Thus, two indirect utility functions, one for each alternative, are estimated:

[1]The rationality test choice sets were removed for the analysis.

$$V_{in} = CONS + \beta_1 ELEM_i + \beta_2 ROAD_i + \beta_3 RSK50_i + \beta_4 RSK75_i + \beta_5 CST1_i + \beta_6 CST2.5_i; \quad (2.6)$$

$$V_{jn} = \beta_7 AGE10_j + \beta_8 AGE40_j + \beta_9 RSK50_j + \beta_{10} RSK75_j + \beta_{11} CST1_j + \beta_{12} CST2.5_j. \quad (2.7)$$

V_{in} represents the indirect utility function of the preventive safety measures alternative and V_{jn} the one of the therapy for a diseased person alternative. In total, there are $n = 210 \cdot (30 - 3) \cdot 2 = 11,340$ individual alternatives in the data set (there are two alternatives in each of the 27 choice sets of which only one could be chosen by the 210 respondents). As the explanatory variables are effects coded, the included constant term CONS in V_{in} reflects the difference in utility that the respondents associate with the labels of V_{in} and V_{jn} (Bech & Gyrd-Hansen, 2005; Salkeld *et al.*, 2000). Hence, the estimated parameter for CONS is the utility that the respondents derive from choosing preventive safety measures over the therapy for a diseased person. A negative sign of CONS can thus be interpreted as empirical support for the presence of an identifiable victim effect. Accordingly, the parameters of the other included regressors are the utility difference between the attribute's level and the corresponding base level. The complete sample of 210 respondents is included in the estimation for this model.

Model 2 is identical to model 1 but for the fact that the cost attribute enters the two indirect utility functions differently. In this model, the cost attribute is treated as alternative generic and enters both functions with its value in EUR million instead of its effects coded representation. Thus, only one parameter is estimated for CST. Although this model is expected to have a slightly worse fit, it has the great advantage that one can readily calculate the marginal rate of substitution (MRS) between CONS and CST. This MRS, multiplied by two because CST reflects only the cost for a risk reduction by 50 percentage points, can then be interpreted to be an estimate of the value that the respondents associate, ceteris paribus, with saving an identified life over a statistical life. Tsuge *et al.* (2005) use a very similar approach to infer the VSL from a DCE by calculating the MRS between a mortality risk related attribute and a cost attribute.

Models 3 and 4 use exactly the same specification as model 1 but they are performed on subgroups of the total sample of 210 respondents. Model 3 only includes respondents who were older than 50 years at the time of the experiment and model 4 includes only respondents who were younger than 50 years. Using the age of 50 years cuts the total sample into two approximately equal parts: There are 99 respondents in model 3 and 111 in model 4. All models were estimated using the statistical package STATA/SE 11.0.

Table 2.5: Perceived difficulties

	Percentage
Difficulty comprehending task	
Not hard at all	9.05 %
Not very hard	54.76 %
Hard	30.00 %
Very hard	6.19 %
Difficulty taking decisions	
Not hard at all	3.81 %
Not very hard	30.48 %
Hard	51.90 %
Very hard	13.81 %

N = 210.

2.4 Results

The respondents understood their task reasonably well. 64% found the task not very hard or not hard at all to comprehend. 30% found it hard and only 6% very hard to understand. Yet, 65% found it hard or very hard to make decisions in the choice sets and only 35% found it not very hard or not hard at all. This suggests that the selected attributes with their levels and the two labels were relevant for the respondents and induced a cognitively demanding trade-off task. In table 2.5, these results are shown in detail.

Table 2.6 reports the results of the four MNL models. In the complete effects coded model 1, all regressors apart from RSK75 and CST2.5 in the 'therapy for a diseased person' alternative equation are significant at the 5% level or lower. CONS is found to be significantly negative, which implies that the respondents associate a lower level of utility with choosing the safety measures alternative than with choosing the therapy alternative. If one agrees that the two labels capture the differences between an individual and a statistical life, then this finding can be interpreted as support for Schelling's assertion that people are more willing to contribute to an identified life than to a statistical life. So, the results of model 1 indicate that the identifiable victim effect exists even when opportunity cost are made explicit to respondents.

In addition, the significant and positive parameters for AGE10 and AGE40 suggest that the respondents associate a higher utility with reducing the mortality risk of younger persons than that of older persons. The sizes of the parameters indicate that this effect of age in terms of years is close to linear. Also, the significant and positive parameter for ELEM is in line with this finding. It means that the respondents appear to prefer an alternative including ELEM over an alternative including NURS, which can be interpreted analogously to the age coefficients. Interestingly, the coefficient of ROAD is significantly negative. One reason for this might be that alternatives aimed at improving the safety of road junctions convey less personalizing information about the people at risk than

alternatives that are targeted at improving safety in nursing homes. The cost parameters have the expected signs and EUR 1 million alternatives are preferred to EUR 4 million alternatives for both safety measures and therapy alternatives. However, for the therapy alternative the respondents do not associate a significant difference in utility between an alternative that costs EUR 2.5 million and an alternative that costs EUR 4 million. Furthermore, the respondents prefer choosing alternatives with a mortality risk of 100% over alternatives with a level of only 50%. The parameter for the 75% risk level, however, is remarkable: For the safety measures alternative this parameter is significant and positive. This implies that the respondents seem to derive a higher utility from reducing a mortality risk of 75% than from reducing a mortality risk of 100% for these alternatives.

The estimated parameters in model 2 are nearly identical to their counterparts in model 1 with regard to size and significance. The model fit is comparable to model 1 with a Pseudo-R^2 of 0.096, but significantly worse according to a likelihood ratio test. From the estimated coefficients for CONS and CST in model 2, one can now simply calculate the MRS between these two attributes by division. To estimate the difference in WTP between saving an identified and a statistical life in EUR, the MRS has to be multiplied by two because CST reports the cost for decreasing the mortality risk by 50 percentage points:

$$WTP = MRS_{CST}^{CONS} \cdot 2 = \frac{-0.234}{-0.169} \cdot 2 \approx EUR\ 2.9\ million. \tag{2.8}$$

Thus, the difference in the WTP between saving an identified and a statistical life that is to die with certainty during the next five years is estimated to be approximately EUR 2.9 million or USD 3.8 million.

However, the results of models 3 and 4 show that the identifiable victim effect in the presence of explicitly stated opportunity cost appears to be dependent on the age of the respondents. In model 3, which includes only respondents above the age of 50, the coefficient for CONS has the same direction as in the models 1 and 2 but has a larger absolute value. In model 4, which includes only respondents below the age of 50, the sign of CONS is significant and positive. This implies that while middle to old aged people seem to have a strong preference for saving an identified life over a statistical life, this preference appears to reverse for younger people. Thus, the confirmation of the identifiable victim effect found in models 1 and 2 is strongly driven by the older part of the respondents. Subgroup analyses with respect to other socio-economic characteristics of the respondents, such as household income, sex, work status or education, turned out not to considerably change the results obtained in model 1. Further notable discrepancies between model 3 and 4 are the sizes and significance levels of the CST and RSK attributes. It seems that the young subgroup in model 4 attaches more importance to the levels of the CST attribute than the middle to old aged subgroup, which is indicated by larger values of CST1. Also, the young subgroup appears to associate a higher level of utility with reducing the high mortality risks.

Table 2.6: Estimated coefficients of the four MNL regression models

Alternative		Model 1	Model 2	Model 3	Model 4
Safety measures alt.	CONS	-0.235***	-0.234***	-0.621***	0.122***
		(0.029)	(0.029)	(0.043)	(0.040)
	ELEM	0.461***	0.463***	0.418***	0.534***
		(0.040)	(0.040)	(0.059)	(0.058))
	ROAD	-0.082**	-0.082**	-0.152**	-0.025
		(0.040)	(0.040)	(0.060)	(0.057)
	RSK50	-0.177***	-0.177***	-0.146**	-0.213***
		(0.041)	(0.041)	(0.061)	(0.058)
	RSK75	0.087**	0.090**	0.066	0.119**
		(0.040)	(0.040)	(0.060)	(0.057)
	CST1	0.150***	-	0.065	0.243***
		(0.040)		(0.060)	(0.057)
	CST2.5	0.086**	-	0.113*	0.069
		(0.040)		(0.059)	(0.057))
Therapy alt.	AGE10	0.511***	0.512***	0.404***	0.646***
		(0.041)	(0.041)	(0.062)	(0.057)
	AGE40	0.294***	0.293***	0.274***	0.346***
		(0.040)	(0.040)	(0.061)	(0.056)
	RSK50	-0.100**	-0.100**	-0.042	-0.100***
		(0.040)	(0.040)	(0.060)	(0.056)
	RSK75	-0.053	-0.051	-0.070	-0.038
		(0.041)	(0.041)	(0.060)	(0.058)
	CST1	0.296***	-	0.227***	0.378***
		(0.041)		(0.061)	(0.058)
	CST2.5	0.027	-	0.104*	-0.043
		(0.040)		(0.060)	(0.057)
Both alternatives	CST	-	-0.169***	-	-
			(0.017)		
	N	210	210	99	111
	n	11340	11340	5346	5994
	χ^2	768.452***	757.34***	446.27***	525.66***
	Pseudo-R^2	0.098	0.096	0.120	0.126

* Significant at 10%, ** 5%, *** 1%. Robust standard errors in parentheses. Model 1: CST alternative specific and effects coded. Model 2: CST alternative generic and linear. Model 3: people ≥50 years. Model 4: people <50 years.

The reported χ^2 statistics in table 2.6 show that all four models are significant at the 1% level according to likelihood ratio tests. Nevertheless, the values of McFadden's Pseudo R^2 are in the range of 0.096 to 0.126, which indicates that the models have relatively poor fits for discrete choice models. However, figure 2.2 shows for all four models that the exhibited choice patterns for the 27 binary choice sets are predicted reasonably well, on average, by the estimated probabilities obtained from the DCE. The figure further illustrates that the respondents appear not to be strongly biased toward either the therapy alternative or the safety measures alternative, as both the predicted probabilities for choosing the therapy alternative and the average percentage of the respondents who actually chose the therapy alternative oscillate around the 50% mark. The slight upward shift of the curves in model 3 compared to model 4 illustrates the middle to old aged subgroup's higher preference for the therapy alternative. In this figure, the solid lines depict the actual percentage of respondents who chose the therapy alternative and the dotted lines depict the estimated probability that a respondent would choose the therapy alternative.

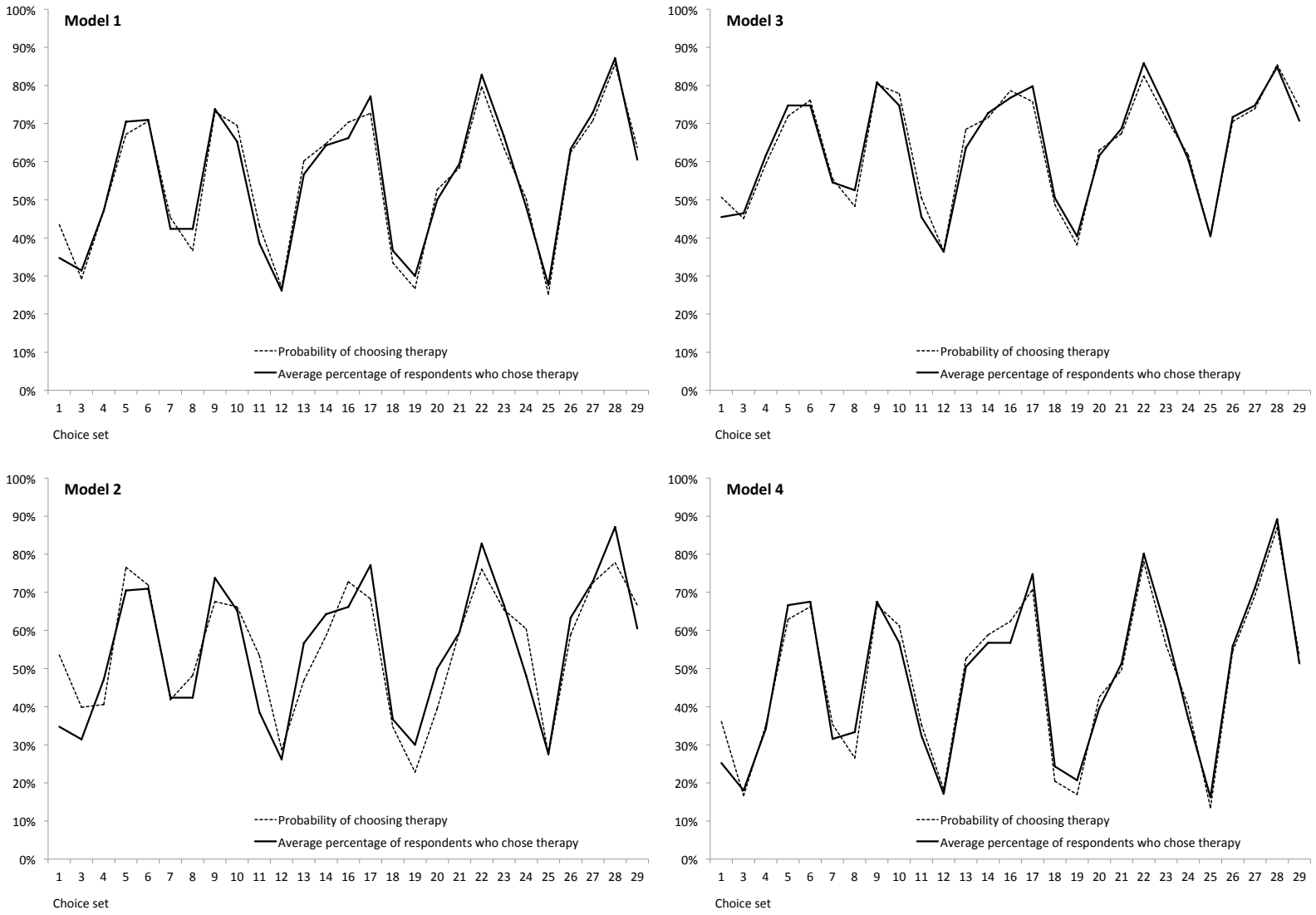

Figure 2.2: Estimated probabilities and actual percentages for choosing the therapy alternative per choice set

2.5 Discussion

This is the first study, to the author's knowledge, that uses a DCE to investigate the identifiable victim effect. Using a DCE has the major advantage that the respondents can be forced to decide between helping either an identified life or a statistical life. This increases the respondents' awareness of the opportunity cost that are associated with their choice. Within the DCE framework two labeled alternatives were generated to accommodate the results of three empirical studies in the relevant literature. By using effects coding and including an alternative specific constant in the assumed underlying utility functions, the results of a MNL model including the full sample show that the respondents associate a higher level of utility with reducing the mortality risk of an identified life than of a statistical life. It is inferred from the DCE estimates that the respondents prefer saving an identified life over a comparable statistical life by an amount of approximately USD 3.8 million in monetary terms. However, a subgroup analysis shows that this result is dependent on the age of the respondents. In contrast to people above the age of 50, younger people appear to prefer saving a statistical life over saving a comparable identified life. Thus, this study provides support for Schelling's assertion and the existence of an identifiable victim effect only on an aggregate level but not for the young subgroup.

The result that respondents above the age of 50 prefer the therapy alternative over the safety measures alternative could be explained by general utility theory. In accordance with Olsen & Donaldson (1998), who find that age appears to have a negative effect on WTP for a helicopter ambulance program but a positive effect on WTP for coronary artery bypass operations and hip replacements, one might argue that the older respondents' more immediate needs influenced their choice behavior. Analogously, one could argue that preventive safety measures were perceived to be of greater value by younger people because they are more likely to benefit from them. However, as the respondents took a decision for a life that is unrelated to their own, it seems unlikely that this notion fully explains the observed divergence in choice behavior between young and middle to old aged people. The observation that the older subgroup appears to be less sensitive to cost may be due to their higher average income. However, Corso *et al.* (2002) find that even if one controls for income, age seems to be positively correlated with WTP for health care programs.

The findings of this study concerning the effect of the age of the persons at risk on choice probability are more in line with intuitive expectations: The older the victims, the less utility the respondents associate with reducing their mortality risk. However, it is important to mention in this context that, as Green & Gerard (2009) point out, the literature is generally inconclusive regarding the importance of age in priority setting scenarios. Thus, one is cautioned not to draw final conclusions from this finding in isolation.

With regard to the reported WTP difference of USD 3.8 million between saving an identified and a statistical life, it is noteworthy that this value is necessarily very sensitive to the chosen attribute levels of the cost attribute. Although some empirical studies suggest

that WTP values elicited through DCEs have a solid degree of external validity (Ryan, 2004; Telser & Zweifel, 2007), the evidence is limited, especially in health related contexts (Ryan & Watson, 2009).

So far, the identifiable victim effect has not played an explicit role in public policy making. However, its existence may be relevant for the current debate on applying distributional weights to the concept of the quality adjusted life year (QALY). Lancsar *et al.* (2011) note that "there is a substantial body of evidence from surveys of the general public implying that, in their evaluation processes, governments and health technology assessment (HTA) agencies should explicitly weight QALYs (or lives) according to contextual factors reflecting characteristics of beneficiaries". According to the findings of the identifiable victim effect literature, the identifiability of the beneficiaries could be one such characteristic. This study shows how a DCE can be used to elicit WTP estimates for the difference between saving an identified life and a statistical life for the context of such debates.

Nevertheless, the normative status of the identifiable victim effect remains obscure. It is questionable whether we should be more compassionate for statistical lives (Lamm, 2001) and acknowledge that identifiability does not appear to be a morally relevant ground for discrimination (Richardson & McKie, 2003) or whether we should succumb to the intuition that – according to empirical and exemplary evidence – we are willing to pay more to help identified lives.

3 The dead-anyway effect from a societal perspective

3.1 Introduction

Over the past forty years a large number of empirical studies investigated people's individual willingness to pay (WTP) for reductions in mortality risk (Viscusi & Aldy, 2003). The results of these studies were often used to infer the value of a statistical life (VSL), which is deemed an important factor in the optimization of public spending in areas such as health care, transportation and environment. Often, an analysis of the observed relationship between wage and risk of jobs in various industries constitutes the basis for these calculations, where it is typically assumed that this relation is linear (see e.g. Liu *et al.*, 1997; Mrozek & Taylor, 2002). However, it has been argued that the elicited WTP values depend on the initial risk levels of the according respondents. In particular, Pratt & Zeckhauser (1996) claim that the individual WTP for a reduction in mortality risk is increasing in initial risk and term their assertion the dead-anyway effect (DAE). More intuitively, the DAE states that the marginal utility of a dollar in the state of death is smaller than in the state of survival. As a consequence of their theoretical analysis, Pratt & Zeckhauser (1996) suggest that, from a normative perspective, individual WTPs should be corrected for the DAE to inform public decisions.

However, empirical evidence on the DAE's existence is inconclusive (Bellavance *et al.*, 2009). Theoretically, the mixed results may be explained by variations in individual characteristics of the respondents. In particular, it is argued that the DAE hypothesis critically depends on assumptions regarding the existence of a bequest motive, the level of human capital investment and the ability to invest in safety-improving expenditures (Breyer & Felder, 2005; Liu & Nelson, 2006). Yet, there are examples of public decisions for which the DAE provides an economic rationale. For instance, the United Kingdom's National Institute for Health and Clinical Excellence (NICE) reported that its Appraisal Committees will consider giving greater weight to quality adjusted life years (QALYs) achieved in the later stages of terminal diseases when appraising end of life treatments (NICE, 2009). Thus, NICE considers accepting higher cost per QALY thresholds regarding treatments for patients with low survival probabilities. This stands in direct contrast with the normative argument that society should not spend more to reduce some risks than others (Pratt & Zeckhauser, 1996).

'The dead-anyway effect from a societal perspective' is joint work with Stefan Felder.

Against this background, we empirically investigate whether people would agree to accommodate individual DAE preferences on the societal level, as it is implied by NICE's decision; i.e. we analyze whether the WTP for a fixed level of risk reduction is increasing in the beneficiaries' initial risk when society is spending the resources. In order to analyze this research question, we use a DCE. This approach also allows us to identify the effect of selected respondent characteristics in this regard. To our knowledge, this is the first study that employs choice analysis to investigate the relationship between initial risk and WTP for risk reduction.

3.2 Methods

3.2.1 The discrete choice experiment

We draw our data from a labeled DCE that was conducted to analyze the difference in WTP between an identified and a statistical life (Sossong, 2012). It featured an orthogonal design with 27 binary choice sets which were supplemented by three rationality test choice sets. Each of the 27 choice sets included an alternative that was labeled 'preventive safety measures' and an alternative that was labeled 'therapy for a diseased person'. The respondents were informed that each alternative aimed at reducing the mortality risk of exactly one person for the next five years by 50 percentage points and that they had to choose one alternative in each choice set. For both alternatives the respondents received information on the 'mortality risk of one person' (RSK) as well as the 'cost for decreasing the mortality risk by 50 percentage points' (CST). RSK could take the values 50%, 75% and 100% and CST comprised the levels EUR 1, 2.5 and 4 million. In addition, the DCE included two further attributes, for which we control in our model specifications. First, the 'area of investment' (INV) was provided for the preventive alternative with the possible levels preventive safety measures for elementary schools ($ELEM$), road junctions ($ROAD$) and nursing homes ($NURS$). Second, the 'age of the diseased person' (AGE) comprising the levels 10-, 40- and 70-year-old person was given for the therapy alternative.

A private market research company identified 599 participants in their German online panel using quota sampling with regard to sex, age, federal state of residency and level of education and invited them to participate in the DCE, which was published in the World Wide Web. The response rate was 61% but further respondents had to be excluded because they either experienced technical difficulties while filling out the DCE, did not pass all three rationality test choice sets or did not provide information on all relevant socio-economic characteristics. The final sample includes 166 respondents.

3.2.2 Model specifications

Drawing on RUT, we use the conventional assumption that the utility of the ith alternative for the nth individual U_{in} includes an observable and an unobservable source of utility.

However, we also account for preference heterogeneity across respondents as proposed by Revelt & Train (1998) and include η_n, a multivariate normal distributed variable with a diagonal variance matrix that captures the nth respondent's deviations from the mean, to depart from the independence of irrelevant alternatives assumption:

$$U_{in} = (\beta + \eta_n)x_{in} + \varepsilon_{in}, \tag{3.1}$$

where x_{in} is a vector comprising all alternative specific attribute-level combinations and interaction effects between levels and socio-economic characteristics of respondent n. We define the observable source of utility $V_{in} = \beta_n x_{in} = (\beta + \eta_n)x_{in}$ and investigate the relationship between initial mortality risk and WTP by estimating three specifications that differ with regard to the structure of V_{in}. In specification 1 the attributes RSK and CST enter in their quantitative representations and in an alternative generic manner whereas the control variables for the attribute levels of INV and AGE, z_i, enter V_{in} effects coded.[1] We also include an alternative specific constant, α_i, to account for the two labels. Thus, specification 1 can be written as

$$V_{in} = \alpha_i + \beta_{1,n}RSK_i + \beta_{2,n}CST_i + \gamma_n z_i. \tag{3.2}$$

Note that RSK, CST and all controls in z are treated as random, i.e. all alternative specific variables are assumed to exhibit preference heterogeneity, which is indicated by the index n of the according βs and γ. In specification 2 we additionally include the interaction effects between respondent income and initial risk, $INC \times RSK$, as well as between respondent age and initial risk, $AGE \times RSK$, to evaluate these characteristics' effects on choice probability:[2]

$$\begin{aligned} V_{in} =& \alpha_i + \beta_{1,n}RSK_i + \beta_{2,n}CST_i + \beta_3 INC \times RSK_{in} \\ &+ \beta_4 AGE \times RSK_{in} + \gamma_n z_i. \end{aligned} \tag{3.3}$$

Finally, specification 3 includes RSK and CST in their effects coded representations as well as the resulting four interaction effects:

$$\begin{aligned} V_{in} =& \alpha_i + \beta_{1,n}RSK50_i + \beta_{2,n}RSK75_i + \beta_{3,n}CST1_i \\ &+ \beta_{4,n}CST2.5_i + \beta_5 INC \times RSK50_{in} + \beta_6 INC \times RSK75_{in} \\ &+ \beta_7 AGE \times RSK50_{in} + \beta_8 AGE \times RSK75_{in} + \gamma_n z_i. \end{aligned} \tag{3.4}$$

[1] See Sossong (2012) for a detailed description of all other attributes and levels. Note that we only include alternative specific variables as controls and no interaction effects with socio-economic characteristics, which is indicated by the sole index i of z.

[2] These interaction effects are treated as non-random because they contain information on the characteristics of the respondents.

In accordance with Revelt & Train (1998) we estimate the following mixed logit (MIXL) model using the add-in module for the statistical package STATA/SE 11.0 provided by Hole (2007) with 500 deterministic Halton draws:

$$Prob_{in}(i|\theta) = \int \frac{exp(V_{in})}{\sum_{j=1}^{J} exp(V_{jn})} f(\beta_n, \gamma_n|\theta) d(\beta_n, \gamma_n) \qquad \forall j \in j = 1, ..., J; i \neq j, \tag{3.5}$$

where $f(\beta_n, \gamma_n|\theta)$ are the density functions of the vectors of utility parameters for the variables in equations 3.2 to 3.4 and θ are the estimated standard deviations and means of their distributions.

3.3 Empirical results

The MIXL results are reported in tables 3.1 to 3.3. It can be inferred from the RSK coefficient in specification 1 that the respondents' utility for a given reduction in mortality risk increases with the initial risk of the beneficiary. However, the significant standard deviation of the estimated RSK coefficient suggests that there is preference heterogeneity regarding this result. In order to interpret this heterogeneity, we relate the magnitudes of the standard deviations to the mean coefficients with $\Phi(-\hat{\beta}_k/\hat{\sigma_k})$, where Φ is the cumulative standard normal distribution, $\hat{\beta}_k$ is the estimated mean coefficient and $\hat{\sigma_k}$ is the estimated standard deviation. According to this approach, 38% of the sample prefer reducing the risk of beneficiaries with low initial risk in specification 1.

Table 3.1: MIXL results of specification 1

	$\hat{\beta}_k$	$\hat{\sigma_k}$	$\Phi(-\hat{\beta}_k \ / \ \hat{\sigma_k})$
RSK	0.700***	2.277***	37.9%
CST	-0.196***	0.005	-
LR χ^2	2162.60***		

N = 166. * Significant at 10%, ** 5%, *** 1%. Controls omitted.

Furthermore, the estimated CST coefficient is in line with the expectation that the respondents prefer lower cost alternatives and there is no preference heterogeneity in this regard. We can calculate the MRS between initial risk and cost to infer the WTP for initial risk reductions on a societal level as follows:

$$WTP = MRS_{CST}^{RSK} = -\frac{0.700 \times 100}{-0.196 \times 1,000,000} \approx EUR\ 36,000. \tag{3.6}$$

Thus, the results of specification 1 suggest that the respondents are willing to increase public spending by EUR 36,000 for an increase of the initial risk of the beneficiary by 1%.[3] This finding supports the conjecture that people agree to accommodate individual DAE preferences when society pays for risk reductions although there is considerable preference heterogeneity in the sample.

Table 3.2: MIXL results of specification 2

	$\hat{\beta}_k$	$\hat{\sigma_k}$	$\Phi(-\hat{\beta}_k \ / \ \hat{\sigma_k})$
RSK	1.562***	2.201***	23.9%
CST	-0.196***	0.013	-
$INC \times RSK$	0.318**	fixed	fixed
$AGE \times RSK$	-0.027*	fixed	fixed
LR χ^2	2155.20***		

N = 166. * Significant at 10%, ** 5%, *** 1%. Controls omitted.

The results of specification 2 indicate that the probability of exhibiting preferences according to the DAE hypothesis on a societal level is increasing in income and decreasing in age. The former result is in line with other empirical studies and theoretical expectations (Bellavance *et al.*, 2009; Pratt & Zeckhauser, 1996). We provide two reasons why the utility derived from reducing high initial risks is decreasing with age. First, it has been shown that age is correlated with a bequest motive which works in the opposite direction of the DAE (Breyer & Felder, 2005; Juerges, 2001). Second, older respondents may be more risk averse and thus draw more utility from reducing the initial risk from 50% to zero than reducing it from 100% to 50%.

In specification 3 the $RSK75$ parameter is not significant. This indicates that the respondents do not associate a significant difference in utility between the initial risk levels 75% and 100%. This implies that the positive relation between WTP for a given risk reduction and initial risk appears to be driven by large increases of initial risk that exceed the 25 percentage points mark.

3.4 Conclusion

Using a DCE we find empirical evidence for a positive relation between the WTP for a given level of risk reduction and initial mortality risk when society spends the resources. In addition, we find that the WTP for a fixed risk reduction does not seem to increase incrementally with initial risk. In fact, our results suggest that increases of the initial risk

[3] We multiply by 100 in the nominator because RSK is denoted in decimals and by 1,000,000 in the denominator because CST is denoted in EUR million in the data.

Table 3.3: MIXL results of specification 3

	$\hat{\beta}_k$	$\hat{\sigma_k}$	$\Phi(-\hat{\beta}_k \ / \ \hat{\sigma_k})$
RSK50	-0.358***	0.549***	25.7%
RSK75	0.011	0.016	-
CST1	0.250***	0.052	-
CST2.5	0.084**	0.053	-
$INC \times RSK50$	-0.853**	fixed	fixed
$AGE \times RSK50$	0.006*	fixed	fixed
LR χ^2	2701.49***		

N = 166. * Significant at 10%, ** 5%, *** 1%. Controls omitted.

below 25 percentage points are not considered significant. The latter result as well as our finding that age promotes DAE behavior casts doubt on the practice of evaluating the effects of initial risk on the VSL solely based on the observed (and assumed to be linear) relationship between wage and (mortality) risk across jobs in different industries.

4 Evaluating the consequences of rheumatoid arthritis

4.1 Introduction

Patients and non-patients tend to attach different utility values to the same state of health that is characterized by suffering from a particular disease. For instance, Boyd *et al.* (1990) find that patients with colostomies assigned significantly higher utilities to this common outcome of treatment for rectal cancer than healthy individuals did. Similarly, Novella *et al.* (2001) observe that non-patient proxies tended to rate Alzheimer's patients' quality of life lower than the patients themselves. In addition, Hays *et al.* (1995) note that patients suffering from epilepsy reported more positive health perceptions and less seizure distress than proxies. Although there is no consensus in the literature that patients generally tend to attach a higher utility value to their current state of health than healthy individuals do, there is a large number of studies that supports this conjecture (Peeters & Stiggelbout, 2010). However, Pyne *et al.* (2009), for example, find that depressed patients reported lower preference scores for depression health states than the general population.

This divergence between patients' and non-patients' valuations of health states leads to the question whose utility values should be used in cost-effectiveness analysis (CEA) (Ubel *et al.*, 2000). Intuitively, one would presume that patients are better informed about the consequences of their illness and public authorities should therefore use the patients' utility values in CEA. Fundamentally, every description of a patient's health state will only be able to convey incomplete or biased information to a member of the general public (Llewellyn-Thomas *et al.*, 1984). Contrary to this presumption, it has been argued that society at large should determine which values are to be used and not the patients because, in the end, it is societal resources that are to be allocated (Gold *et al.*, 1999). The debate on whose values should be used in CEA is still ongoing and has been deemed to ultimately be a normative judgment (Brazier, 2008).

With this paper, we follow the suggestion of Ubel *et al.* (2003). According to these authors, "the most important challenge facing researchers [...] is to conduct studies that shed light on why these discrepancies occur." We aim at contributing to gain a better understanding of the differences between patient and non-patient evaluations of health states characterized by typical consequences of rheumatoid arthritis (RA). In particular,

'Evaluating the consequences of rheumatoid arthritis' is joint work with Stefan Felder, Malte Wolff and Klaus Krüger.

we explore whether the two groups attach different utility values to disease specific consequences as opposed to suffering from the disease as a whole. This allows us to identify potential differences in the underlying factors that might add to discrepancies between the two groups' absolute health state evaluations and thus infer more informed policy implications.

Therefore, we conducted a discrete choice experiment (DCE) with a final quasi representative sample of 200 members of the German population and a final sample of 227 RA patients. In the DCE, both patients and non-patients were presented with the same 18 pair-wise comparisons of hypothetical health states characterized by alternating levels of the selected consequences of RA. The conduct of the DCE with a sample of patients was approved by the local ethic committee of the University of Duisburg-Essen in October 2011.

4.2 Rheumatoid arthritis

RA is a chronic inflammatory disease that affects joints. In contrast to many other diseases that are subsumed under the colloquial term rheumatism, RA is an autoimmune disease that becomes more severe over the course of time. Approximately 0.2% to 1.2% of the adult population worldwide is affected by RA whereas its annual incidence rate varies between 25 and 115 cases per 100,000, depending on case definitions and geographical regions. Also, the prevalence rate for females tends to be considerably higher than for males (Carmona *et al.*, 2010).

In many cases, the first symptoms of RA include swollen and stiff finger and wrist joints as well as pain. However, larger joints like the shoulder and knee can also be affected. Additionally, patients often suffer from fatigue and stiffness, especially in the mornings. In the course of the disease, the adjacent cartilages and bones are damaged and, if RA remains untreated, the affected joints are ultimately destroyed.

Due to pain and swollen joints, patients are often very limited in their fine motor skills, so that simple everyday actions, such as unbuttoning a shirt or opening a bottle can become difficult. Employed patients are more frequently issued a statement of incapacity of work than healthy individuals and some even have to stop working (Merkesdal *et al.*, 2001). Although symptoms and functional limitations can decline, inflammatory induced destructions are irreparable. In the long term, many patients need joint replacements or become dependent on care. As a consequence, RA causes significant direct and indirect cost for an economy. In the case of Germany, the additional cost per patient year was estimated to amount to EUR 3,830 of which about 23% were productivity related and caused by sick leave, work disability or other RA-related work loss (Kirchhoff *et al.*, 2011). By performing a simple indicative calculation and assuming that about 800,000 people are affected in Germany, we note that RA inflicted cost of approximately EUR 3 billion on the German economy in 2002.

4.3 Methods

4.3.1 DCE construction

In order to construct a DCE that allows us to investigate whether patients and non-patients associate different utility values with specific consequences of RA we took the following conventional and interdependent steps (Kløjgaard *et al.*, 2011). First, we defined the decision-context for the respondents within the DCE. Second, we identified the patients' most relevant consequences and symptoms of RA which entered the DCE in the form of attributes. Third, we assigned levels to these attributes. Fourth, we generated the experimental design, i.e. the list of choice sets which differ with regard to the levels of the selected attributes. Fifth, we tested the DCE questionnaire with patients. In the following subsections we will briefly explain how we proceeded in each step.

Decision-context

Dolan & Kahneman (2008) argue that the more conventional methods that are used to elicit utility values for health states, namely standard gamble (SG) and time trade-off (TTO), suffer from the deficiency that they are unlikely to generate meaningful utility estimates for health states that are based on different experiences. Their main argument is that while patients adapt to their deteriorated health condition, non-patients tend to underestimate the utility effect of this adaptation process and therefore also tend to underrate the utility associated with patients' health states. To mitigate this point of criticism we presented patients and non-patients with pair-wise comparisons of generic hypothetical health states that were characterized by typical consequences of RA and asked them to decide in each comparison or choice set which of the two health states they consider to be worse according to their personal opinion. As a result, we circumvented the potential need to correct non-patient evaluations by an adaptation factor because we, loosely speaking, included a controllable reference case in each choice set, that is the forgone alternative. Furthermore, we provided a brief description of RA to non-patients, which was similar to the one given in section 4.2.

Attributes

In order to identify the consequences and symptoms of RA that are most relevant to patients, we reviewed the relevant literature and conducted four expert discussion sessions with researchers and practising physicians in the field of RA between November 2010 and June 2011. All experts agreed to support this project by providing advise during the first meeting of AbbVie's (formerly Abbott) health economic expert committee in rheumatology in May 2010.

As a result, we decided to include the following five consequences and symptoms of RA in the study: 'inability to work for three months', 'severe fatigue', 'severe problems with

buttoning a shirt or blouse', 'severe pain' and 'duration of treatment until condition starts improving'. Of course, there exist many more attributes of potential relevance. Although there is no clear guidance on how many attributes a respondent can be presented with without impacting the random component variability within the DCE framework, most studies tend to assume that four to six attributes are acceptable and that beyond the choice tasks become cognitively too complex (Ryan & Gerard, 2003). We followed the assumption of the majority of studies and decided to include a maximum number of five attributes in the DCE. Generally, we closely examined all attributes included in the Arthritis Impact Measurement Scales (AIMS) with regard to fit in our research design (Meenan *et al.*, 1980). During this selection process we also strongly considered the following – but ultimately not included – two attributes: 'problems with walking some hundred meters' as an attribute describing limitations related to the lower body parts and an attribute related to a deteriorated level of social interaction due to RA, e.g. 'inability to attend gatherings with friends or relatives on a regular basis'. However, as a result of our expert discussions, we came to the conclusion that due to modern treatment methods there are only few patients who suffer from serious limitations to mobility due to RA. Thus, including such an attribute would not have been very relevant for a majority of patients in our sample. In addition, we decided to not include a social interaction attribute as we were concerned about serious levels of preference heterogeneity across patients in this respect.

We had a specific research interest in exploring whether there is a discrepancy between patients' and non-patients' utility valuation of being incapacitated for work. First, work loss due to RA is very typical among patients. As a result of a systematic literature review, Burton *et al.* (2006) find that in their sample 66% of employed RA subjects experienced work loss due to RA in the previous 12 months for a median duration of 39 days. Second, it was observed that patients who are incapacitated for work report more pain and depression than patients who work (Fifield *et al.*, 1991). This result might suggest, although the causality does not become clear in the cited study, that remaining employed is important to patients. Thus, we found it worthwhile to investigate the relevance of being employed after the onset of the disease for patients and whether patients' and non-patients' perspectives are conform in this regard. We included the inability to work as a binary attribute, so that we had to choose a time period for which the respondent had to imagine to be incapacitated for work. We chose a period of three months and termed the attribute accordingly.

In addition, we included 'severe fatigue' as an attribute because it is common in RA and reported to be of great relevance for patients. Wolfe *et al.* (1996) find that fatigue was present in 88-98% of their patient sample and 41% of the sample reported clinically important levels of fatigue. Also, numerous studies find that fatigue is an important determinant in the patients' reported quality of life (Campbell *et al.*, 2012; Kirwan & Hewlett, 2007; Minnock *et al.*, 2003; Swain, 2000). By choosing to add the adjective 'severe' we wanted to convey the notion of a clinically relevant level of fatigue to the respondents.

Furthermore, we included the attribute 'severe problems with buttoning a shirt or blouse'

as a proxy to estimate the (negative) utility associated with functional limitations of fine motor skills due to RA. In accordance with the experts' opinions, we deemed it necessary to provide a specific example for a certain type of limitation to increase the level of homogeneity of the respondents' understanding of this attribute. We chose the example of buttoning a shirt or blouse because in their article presenting the AIMS, Meenan *et al.* (1980) found this activity to achieve the median score on a Guttman scale ordering items that describe limitations to dexterity due to RA. The according item was also kept in further developments of the AIMS, in particular the AIMS2 (Meenan *et al.*, 1992) as well as the German adaptation of the AIMS2 short form (AIMS2-SF) (Rosemann *et al.*, 2005), emphasizing its relevance.

'Severe pain' was a natural inclusion in the DCE because it is a consequence of the inflammatory nature of the disease and often reported to be of great relevance by patients according to the experts. According to the literature, pain is viewed as one of the most troublesome features of RA (Covic *et al.*, 2000; Kazis *et al.*, 1983).

The previously mentioned four attributes are the ones of specific research interest. However, as we wanted to be able to compare these attributes on the same basis of interpretable marginal rates of substitution (MRS), we decided to include the comparator attribute 'duration of treatment until condition starts improving'. Kløjgaard *et al.* (2011) use a very similar attribute, which they label 'waiting time for the given treatment effect to occur', as a payment vehicle in their DCE on different treatments for spine surgery.

Levels

As pointed out, we included the 'inability to work for three months' as a binary attribute. Accordingly, the two levels associated with this attribute were labeled 'yes' and 'no'. With regard to the attributes 'severe fatigue', 'problems with buttoning a shirt or blouse' and 'severe pain' we chose levels that are close to the AIMS2-SF questionnaire's levels because it is a validated and commonly used instrument. In the AIMS2-SF respondents are asked to indicate whether they had been affected by the according question or item during the past four weeks either 'all days', 'most days', 'some days', 'few days' or 'no days'. Since our respondents had to decide between two competing hypothetical health states, we generalized the time frame from 'during the past four weeks' to 'per month'. Moreover, as we wanted to be able to have the attributes enter our model specifications in a quantitative representation, that is to say with their value, and calculate readily interpretable MRS, we removed some ambiguity from the AIMS2-SF levels by providing concrete numbers of days. Thus, we decided on the following three levels for these attributes: 'all days per month', 'about 15 days per month' and 'about 5 days per month'. We kept the word 'about' to draw a more realistic picture of the health states for the respondents. Additionally, we decided to include three levels because that allowed us to estimate non-linear effects in comparison to just two levels. For the comparator attribute 'duration of treatment until conditions starts improving' we chose the levels '1 month', '3 months' and '6 months', which were deemed to be realistic time spans according to the experts

we discussed this issue with. Table 4.1 provides an overview of all attributes with their according levels.

Table 4.1: Attributes and levels

Attribute	Levels
Inability to work for three months (JOB)	yes (YES) no (NO)
Severe fatigue due to RA (FAT)	all days per month (FAT30) about 15 days per month (FAT15) about 5 days per month (FAT5)
Severe problems with buttoning a shirt or blouse (FMS)	all days per month (FMS30) about 15 days per month (FMS15) about 5 days per month (FMS5)
Severe pain due to RA (PN)	all days per month (PN30) about 15 days per month (PN15) about 5 days per month (PN5)
Duration of treatment until condition starts improving (DUR)	1 month (DUR1) 3 months (DUR3) 6 months (DUR6)

Experimental design

A full-factorial design, i.e. a design that incorporates all possible combinations of attributes and levels in pairwise comparisons, would include 162 ($2^1 \times 3^4$) choice sets.[1] To reduce this number of choice sets to a manageable amount for the respondents we first considered generating a design that maximizes d-efficiency. A d-efficient design minimizes the standard errors of the estimated parameters in a multinomial logit (MNL) model in comparison to other feasible designs and thus increases the likelihood of estimating significant parameters (Kuhfeld, 2010). While it is of course desirable to minimize the parameter estimate variances, which are the diagonal elements of $C^{-1} = (X'X)^{-1}$, where X is the design matrix and $C = (X'X)$ is the information matrix, statistical efficiency has to be balanced with response efficiency. D-optimal or nearly d-optimal designs do not per se exclude implausible comparisons of alternatives or take into account possible cognitive limitations of the respondents with regard to the complexity of the decisions at hand (Johnson *et al.*, 2013). Due to the fact that we expected many of our respondents to be in a condition that will not allow us to present very complex trade-off situations,

[1] Note that we employ a generic DCE. That means that the alternatives in our choice sets are not labeled, at least not in a meaningful way.

we opted to exercise more influence on the final set of choice sets and took the following approach, deliberately sacrificing statistical efficiency for response efficiency and taking into account that the variances of our parameter estimates may become too large to yield significant results.

We generated an orthogonal main effects plan (OMEP), assuming that higher-order interaction effects are equal to zero, which is an acceptable assumption in most cases as it was found that over 80% of the preference structure is explained by main effects and there is no specific reason to believe that interaction effects play a considerable role in the research context of this study (Emery & Barron, 1979; Louviere, 1988; Permain *et al.*, 1991). As a consequence, the majority of studies in health economics chooses to generate designs that do not include orthogonal interaction effects to keep the number of choice sets acceptably small for the respondents (Ryan & Gerard, 2003). The resulting OMEP included 16 profiles. We then proceeded to randomly pair these 16 profiles with copies of themselves, making sure that a copy is not paired with its original and identical counterpart, to construct 16 choice sets. Subsequently, we made changes to the design to achieve a plausible and not overly complex set of decisions by manually re-pairing some of the alternatives. It is noteworthy that by doing so we retained within-alternative-orthogonality of the attributes because our DCE included generic alternatives that were not meaningfully labeled (Hensher *et al.*, 2007). However, as we already noted, our resulting design was not d-optimal or near d-optimal since the according variance-covariance matrix of our parameter estimates were not diagonal. This resulted from the fact that our manual re-pairing did not ensure that the design was balanced (Johnson *et al.*, 2013).

To ensure that our design was still capable of estimating all parameters of interest, we calculated its d-efficiency following the approach of Street *et al.* (2005):

$$D - efficiency = \left[\frac{det(C)}{det(C_{optimal})}\right]^{\frac{1}{p}}, \tag{4.1}$$

where $det(C)$ is the determinant of the information matrix C, p is the sum of all parameters to be estimated and $det(C_{optimal})$ is the upper bound for the determinant of the information matrix for estimating main effects and calculated for this specific design as follows.

$$det(C_{optimal}) = \prod_{q=1}^{k} \left(\frac{2}{m^2(l_q - 1)\prod_{i=1, i\neq q}^{k} l_i}\right)^{l_q - 1}, \tag{4.2}$$

where k is the number of attributes, m is the number of alternatives in each choice set and l_i is the number of levels of attribute i.

First, we calculated the information matrix $C = B\Lambda B'$, where B is the matrix of contrasts for the effects to be estimated and Λ is the matrix of second derivatives of the likelihood function. As a second step, we calculated $det(C)$ as well as $det(C_{optimal})$ and found that

the d-efficiency of our design equals 21%. Although this is a rather low efficiency value, it is larger than zero, so that we can estimate all parameters of interest with the according design. We decided to conduct the DCE with this statistically inefficient design for mainly two reasons. First, we were very satisfied with its response efficiency, in particular with its plausibility and complexity of choice tasks, which was emphasized by comments of test respondents. Second, we were willing to accept higher standard errors of parameter estimates because we were not so much interested in the precision of parameter estimates because we, for instance, did want to estimate specific willingness to pay (WTP) values; we were rather interested in determining the significance of the attribute levels' estimates and thus did not mind a more conservative or inferior design with regard to the ability to produce significant results.

In addition to the 16 choice tasks, we included two rationality tests that comprised one dominant alternative. We did this to be able to filter out all respondents that showed irrational behavior by selecting the inferior alternative in these choice sets. This kind of test is in essence a test for non-satiation (Lancsar & Louviere, 2006; Miguel *et al.*, 2005). Table 4.2 shows the final list of all choice tasks.

Table 4.2: List of all choice sets in experimental design

Choice	State A					State B				
1	NO	FAT30	FMS15	PN30	DUR1	YES	FAT5	FMS30	PN30	DUR6
2	Rationality test[a]									
3	NO	FAT15	FMS5	PN30	DUR1	YES	FAT30	FMS30	PN15	DUR1
4	NO	FAT15	FMS30	PN5	DUR3	YES	FAT5	FMS5	PN30	DUR3
5	YES	FAT5	FMS5	PN5	DUR1	NO	FAT30	FMS5	PN5	DUR6
6	YES	FAT15	FMS15	PN5	DUR6	NO	FAT5	FMS5	PN5	DUR1
7	NO	FAT5	FMS30	PN5	DUR1	NO	FAT5	FMS5	PN15	DUR6
8	NO	FAT5	FMS5	PN5	DUR1	YES	FAT30	FMS5	PN5	DUR3
9	YES	FAT5	FMS5	PN30	DUR3	NO	FAT5	FMS15	PN15	DUR3
10	YES	FAT5	FMS15	PN5	DUR1	YES	FAT15	FMS5	PN15	DUR1
11	NO	FAT30	FMS5	PN5	DUR6	NO	FAT30	FMS15	PN30	DUR1
12	YES	FAT5	FMS30	PN30	DUR6	YES	FAT15	FMS15	PN5	DUR6
13	YES	FAT30	FMS30	PN15	DUR1	NO	FAT15	FMS5	PN30	DUR1
14	NO	FAT5	FMS5	PN15	DUR6	YES	FAT5	FMS15	PN5	DUR1
15	YES	FAT15	FMS5	PN15	DUR1	NO	FAT5	FMS30	PN5	DUR1
16	YES	FAT30	FMS5	PN5	DUR3	NO	FAT15	FMS30	PN5	DUR3
17	NO	FAT5	FMS15	PN15	DUR3	YES	FAT5	FMS5	PN5	DUR1
18	Rationality test[a]									

[a] One alternative is dominant.

Test of DCE with patients

In December 2011, we conducted tests with 13 patients at Rheumazentrum München. Either before or after seeing the physician, the test patients were asked if they agreed to take part in a study on RA by the doctor's receptionists or the physician himself. In case of agreement, the patients were led into a separate room and first given a brief background on the project. Then they were asked to complete the paper version of the DCE and answer selected socio-economic questions. Afterwards, the interviewer discussed comprehension problems and any remarks with the test patients. On average, an interview lasted about 45 minutes.

The majority of test patients stated that they had no problems at all related to comprehending the task at hand. However, about half of them stated that it was very tedious to go through the choice sets because they contained a large amount of text. Consequently, we decided to include additional visual scales for the attributes FAT, FMS and PN. Each scale had 30 small units that were filled red according to the shown level. For instance, for a health state that included 15 days of severe pain per month, 15 units were filled red. Other than that we concluded from the test that no further changes to the questionnaire were necessary. Figure 4.1 shows an example of a final choice set.

Decision 1	**State A**	**State B**
Inability to work for 3 months	no	yes
Severe fatigue due to RA	all days per month	5 days per month
Severe problems with buttoning a shirt or blouse	15 days per month	all days per month
Severe pain due to RA	all days per month	all days per month
Duration of treatment until condition starts improven	1 month	6 months
I find the following state worse	☐	☐

Figure 4.1: Example of a choice set (original in German)

4.3.2 Sample and data collection

We developed an online questionnaire that included a brief introduction, a brief description of RA for the non-patient version, the DCE and conventional socio-economic questions. From November to December 2011 we published the non-patient DCE online. A

market research company was contracted to provide a quasi-representative sample of the German population. We ensured that we could identify the market research company's panelists by transmitting identification numbers to the company so that we could assign every server access to a specific panelists. Quota sampling with regards to the characteristics sex, age, residency and income was used to identify the respondents of the non-patient sample. In total, 585 non-patient panelists were invited to participate in the study. 165 of these did not react to the invitation and 114 took at least a look at the questionnaire but did not complete it. Of the remaining 306 panelists who completed the DCE we excluded 60 subjects because they did not pass the two rationality tests. In addition, we excluded further 11 respondents because they answered the whole DCE questionnaire within five minutes. We assumed that this time frame does not suffice to answer the DCE in a serious manner. Finally, we had to exclude further 35 subjects since they encountered technical difficulties while completing the DCE and parts of the data were lost. In the end, exactly 200 respondents were left in the non-patient sample.

In order to collect data from patients, we took two different approaches. First, we visited six doctor's offices from June to August 2012. Due to financial constraints, we could not cover doctor's offices in many geographical regions of Germany but had a focus on the western part of the country. Most of the physicians allowed us to approach patients by ourselves and ask them if they agreed to participate in the study. 96 of the patients who completed the paper version of the DCE in the doctor's offices passed the two rationality tests and allowed us to include their answers in this study.

Second, we recruited 137 patients via "Rheumaliga", the German patient support group for RA. The editors of the support group's member's journal agreed to place a notification about the study, including a link to the online version of DCE, in the journal's edition of August 2012. We clearly indicated in the journal notification as well as on the first page of the online version of the DCE that only patients with RA were asked to complete the DCE. However, we had no possibility to ensure that only patients participated in the online survey. Yet, as we recorded the timestamps of server accesses, we found that the vast majority of respondents accessed the online DCE within a few days after the publishing date of the journal's edition. We take this as an indicator that the majority of people who completed the DCE in August 2012 were indeed patients.

Consequently, the final sample of patients and non-patients includes 427 respondents. Table 4.3 lists selected descriptive statistics of the sample for patients and non-patients. There are major differences between the two groups. First, the percentage of female respondents is about 30 percentage points higher in the patient sample than in the non-patient sample. This is not surprising since prevalence rates for females are considerably higher than for males. Second, in patient households, there tend to live fewer persons on average. This may be due to the fact that patients are on average five years older than non-patients. Finally, there is a difference of 10 percentage points between patients and non-patients with regard to the proportion of being employed.

Table 4.3: Descriptive statistics of the sample

	Patients $n = 227$	Non-patients $n = 200$
Sex		
Female (%)	81.06%	50.50%
Age		
Mean (yrs)	54.74	49.29
Std. dev.	12.91	14.92
Household income per month		
EUR 0 - <1000	11.45%	5.50%
EUR 1000 - <2000	18.94%	16.50%
EUR 2000 - <3000	19.82%	22.00%
EUR 3000 - <4000	12.33%	13.00%
EUR 4000 - <5000	5.73%	9.00%
>= EUR 5000	9.69%	10.00%
N/A	22.03%	24.00%
Household size		
1 person	22.91%	13.50%
2 persons	51.10%	44.00%
3 or more persons	25.99%	42.50%
School education		
No certificate	0.44%	0.00%
Lower sec. educ.	29.07%	24.50%
Middle school	32.16%	34.50%
A level	34.36%	36.50%
Other	3.52%	2.00%
N/A	0.44%	2.50%
Occupation		
Working	50.66%	60.00%
Not working	44.93%	35.50%
N/A	4.41%	4.50%
Questionnaire		
Online	57.71%	100.00%
Paper	42.29%	0.00%

4.3.3 Data analysis

In order to analyze the data collected from the online and paper surveys, we estimate four MNL models. Model 1 includes only the patient subsample and $\sum_{q=1}^{k}(l_q - 1)$ parameters to be estimated, where k is the number of attributes and l_q is the number of levels of attribute q. We assume a linear-in-parameters indirect utility function V_{in}, describing the utility of alternative i for respondent n, and each level enters V_{in} additively and dummy

coded, where the base level is usually defined to be the worst level of each attribute. For instance, the worst level of 'duration of treatment until condition starts improving' is '6 months' which thus constitutes the base level of this attribute. The only exception to this definition pattern is with regard to the attribute 'inability to work for three months'. Since the preference patterns behind being incapacitated for work are very unclear, we define the base level for this attribute to not be incapacitated for work. Please note that we estimate models that are based on a generic DCE. Thus, using effects coding does not yield any advantage over using dummy coding because it does not make sense to include alternative specific constants in the specifications. Accordingly, the assumed underlying indirect utility function V_{in}, for alternative i and subject n, can be written as follows for model 1:

$$\begin{aligned} V_{in} =& \beta_1 JOB_{YES,i} + \beta_2 FAT5_i + \beta_3 FAT15_i + \beta_4 FMS5_i + \beta_5 FMS15_i \\ &+ \beta_6 PN5_i + \beta_7 PN15_i + \beta_8 DUR1_i + \beta_9 DUR3_i + \epsilon_{in}. \end{aligned} \tag{4.3}$$

Note that ϵ represents random variation across discrete choices and is assumed to be extreme value type 1 distributed, so that we can estimate the probability that alternative i is preferred to any alternative j with the MNL model:

$$Prob_{in} = \frac{exp(V_{in})}{\sum_{j=1}^{J} exp(V_{jn})} \qquad \forall j \in j = 1, ..., J; i \neq j. \tag{4.4}$$

Model 2 is identical to model 1 with regards to the estimated specification. However, it only contains the answers of the non-patient subsample. We believe that it is advantageous to present patient and non-patient parameter estimates separately as a first step because it provides a good intuition for the potential differences between the two groups' preference structure. Nevertheless, the following models incorporate both patients and non-patients to allow statistical inference about the effect of being a patient on the selected consequences of RA.

In model 3 all attributes, apart from 'inability to work for three months', enter the underlying indirect utility function V_{in} with their quantitative representation, so that we estimate only one parameter for the main effect of each attribute. In addition, we include interaction terms between the nth respondent characteristic of being a patient, PAT_n, and all consequences of RA. This approach allows us to test if the parameter estimates of patients and non-patients are statistically different by using Wald tests for the interaction terms. Thus, the specification of model 3 can be written as follows:

$$\begin{aligned} V_{in} =& \beta_1 JOB_{YES,i} + \beta_2 FAT_i + \beta_3 FMS_i + \beta_4 PN_i + \beta_5 DUR_i \\ &+ \beta_6 JOB_{yes,i} \times PAT_n + \beta_7 FAT_i \times PAT_n + \beta_8 FMS_i \times PAT_n \\ &+ \beta_9 PN_i \times PAT_n + \beta_{10} DUR_i \times PAT_n + \epsilon_{in}. \end{aligned} \tag{4.5}$$

The final model 4 is is an extension of model 3. In addition to estimating interaction effects between being a patient and consequences of RA, we include the interaction effects between the 'inability to work for three months' and the following respondent characteristics: household income (INC_n)[2], sex (SEX_n), age (AGE_n), household size (HH_n)[3] and being employed (EMP_n). The specified indirect utility function of model 4 can be written as follows:

$$\begin{aligned} V_{in} =& \beta_1 JOB_{YES,i} + \beta_2 FAT_i + \beta_3 FMS_i + \beta_4 PN_i + \beta_5 DUR_i \\ &+ \beta_6 JOB_{yes,i} \times PAT_n + \beta_7 FAT_i \times PAT_n + \beta_8 FMS_i \times PAT_n \\ &+ \beta_9 PN_i \times PAT_n + \beta_{10} DUR_i \times PAT_n + \beta_{11} JOB_i \times INC_n \\ &+ \beta_{12} JOB_i \times SEX_n + \beta_{13} JOB_i \times AGE_n + \beta_{14} JOB_i \times HH_n \\ &+ \beta_{15} JOB_i \times EMP_n + \epsilon_{in}. \end{aligned} \tag{4.6}$$

4.4 Results

Table 4.4 reports the results of the categorical questions 'How difficult was it for you to understand what you were asked to do?' and 'How difficult was it for you to make decisions in the scenarios?'. 89% of the non-patients and 76% of the patients found it not hard at all or not very hard to understand the task at hand. Only 3% of the patients and none of the non-patients found it very hard to comprehend what they were asked to do. This shows that the respondents understood the choice task acceptably well. However, it is noteworthy, that non-patients seem to have had a better understanding on average. Concerning the reported difficulty of taking decisions in the DCE, the two groups show very similar results. About 67% of the non-patients and 63% of the patients found it not hard at all or not very hard to make decisions in the choice sets. 3%, respectively 6%, found it very hard to make choices. It is typical for DCEs that the process of making a decision is perceived to be harder than understanding the task. This is an indication that the respondents had to consider relevant trade-offs.

We report the results of the four described MNL models in table 4.6. All models are significant on the 1% level according to a likelihood-ratio test which can be inferred from the significance of the reported χ^2 statistics. Also, the reported pseudo-R^2 are between 0.367 and 0.380 which indicate that the models have comparably good fits for DCEs.

With regard to models 1 and 2 we find that all, but DUR3, of the included dummy coded attribute levels significantly affect choice probability. This indicates that, overall, the attributes and levels were considered to be of relevance by patients and non-patients alike. In model 1 and 2 all significant parameters of FAT, FMS, PN and DUR have negative

[2] We include the lower bounds of the income categories in the regression. The results are, however, robust with regard to significance when alternatively including the category means or the upper bounds and excluding responses from the highest and boundless income category.

[3] We include the category '3 or more persons' as 3 in the regression analysis.

Table 4.4: Perceived difficulties of DCE

	Non-patients	Patients
Difficulty comprehending task		
Not hard at all	31.50%	23.79%
Not very hard	57.50 %	52.42%
Hard	11.00%	20.70%
Very hard	0.00%	3.08%
Difficulty taking decisions		
Not hard at all	7.50%	10.57%
Not very hard	59.00%	52.42%
Hard	30.50 %	30.84%
Very hard	3.00%	6.17%

N(Non-patients) = 200. N(Patients) = 227.

signs. This was highly expected because a negative sign indicates that the according level exhibits a negative effect on choice probability compared to the selected base level. Note that the respondents were asked to choose the worse of the two alternatives in each choice set and that the base level was chosen to be the worst level of the mentioned attributes. Thus, loosely speaking, a negative sign of these parameters means that the better the level of the attribute the better the perception of the alternative, which is intuitive. In addition, we would expect that the levels FAT5, FMS5, PN5 and DUR1 are smaller than their according second worst levels FAT15, FMS15, PN15 and DUR3. This is only the case for FAT and PN in the non-patient model and for FAT, FMS and PN in the patient model. In case of FMS, non-patients are not able to distinguish between 'severe problems with buttoning a shirt or blouse' for 15 or 5 days utility-wise, while patients were able to do so. Concerning DUR, both patients and non-patients do not attach significantly different utility values to a duration of 3 or 6 months of treatment until the condition starts improving. According to several comments by patients during the interviews, 3 and 6 months are both perceived to be unacceptably long time periods in this context for them, which may explain the insignificance of DUR3.

Overall, the results of model 1 and 2 are very similar apart from the different evaluation of the attribute 'inability to work for three months' (JOB_{YES}). While being incapacitated for work is associated with a significant negative utility impact by patients[4] the opposite is the case for non-patients. Thus, the valuation of being able to work constitutes a significant discrepancy between patients' and non-patients' assessments of health states that are characterized by consequences of RA.

As a result of model 3 we find that all estimated parameters for the quantitative representations of the attributes have the expected sign and are significantly different from zero: The more days per month (FAT, FMS and PN) or the longer (DUR) one suffers in a

[4]Note that a positive sign means that the according level increases choice probability and the respondents were asked to choose the worse of the two alternatives in every choice set.

health state from a symptom, the higher the probability that the according health state is chosen as the worse of the two alternatives. In addition, we find that among all interaction terms only the one between the 'inability to work for three months' and the respondent being a patient is significant. This observation confirms the intuition gained in models 1 and 2 that the estimated parameters for all attributes but JOB do not significantly differ between patients and non-patients. Thus, JOB is the only relevant discrepancy between the two groups according to model 3.

Model 4 expands model 3 by including additional interaction terms between JOB and selected respondent characteristics. The results show that respondent characteristics related to sex (SEX), age (AGE) and the status of being employed (EMP), are not significantly correlated with the evaluation of the 'inability to work for three months'. However, the higher the income of the respondent's household, the higher is the effect of JOB on choice probability for the according subject ($JOB \times INC$). Thus, a respondent with a high household income will, ceteris paribus, attach a lower utility value to the condition of being incapacitated for work than a a respondent with a low household income. Additionally, we find a significant opposite effect for the respondent's household size ($JOB \times HH$). The larger the household, in terms of persons living in the household, the lower is the effect of JOB on choice probability. Accordingly, a respondent living in a small household will, ceteris paribus, associate a lower utility with being incapacitated for work than a respondent living in a large household.

Furthermore, table 4.5 shows the calculated MRS between the attributes of research interest (JOB, FAT, FMS and PN) and the comparator attribute (DUR) based on the results of model 4.[5] This table can be interpreted as a ranking of the severity of RA consequences, at least for the attributes PN, FAT and FMS. In particular, the respondents are willing to accept 8.59 additional days in treatment until the condition starts improving for reducing the number of days per month on which they suffer from severe pain by one. As the respondents are willing to accept more days in a bad condition for an improvement in pain than, e.g., an improvement in fatigue, it can be inferred that pain is perceived to be more severe than fatigue. However, one must be careful with regard to the interpretation of the inability to work. According to table 4.7, the respondents are willing to endure 20.52 additional days until the condition starts to improve for not being incapacitated for work for three months. Note that as the inability to work for three months is a binary attribute, we cannot infer the severity of an incapacitation for work of just one day's length. This makes it difficult to include the JOB related estimated parameters in a severity ranking. Also note that the different signs of $JOB \times PAT$ and JOB reflect the divergent valuations of patients and non-patients.

[5] Note that this calculation is critically dependent on our assumption regarding the structure of the indirect utility function V_{in} and our assumption about the absence of significant interaction effects among the mentioned attributes.

Table 4.5: MRS between attributes

Attribute	Additional days accepted in bad condition[a]
$JOB \times PAT$	20.52
PN	8.59
FAT	5.33
FMS	2.37
JOB	-23.26

[a] Number of additional days in bad condition the respondent is willing to accept for the reduction of one day of suffering from the according attribute (PN, FAT, FMS) or for not being incapacitated for work for three months ($JOB \times PAT$, JOB). We calculate the shown values by dividing the estimated parameter of the according attribute by the estimated parameter of DUR in model 4 and multiplying the result by 30. We multiply by 30 because DUR is not coded as the number of days but as the number of months.

4.5 Discussion

We use a DCE to explore the discrepancies between patients' and non-patients' evaluations of health states that are characterized by typical consequences of RA. Our results show that the two groups only deviate in their evaluation with regard to being incapacitated for work for three months and the ability to differentiate between a limitation of fine motor skills for 5 or 15 days per month. For all other consequences of RA we observe nearly congruent utility valuations. Furthermore, we find that the respondent's household income and household size are correlated with the perceived severity of incapacitation for work in addition to the patient status.

There exist several studies that provide possible explanation approaches for discrepancies between patient and non-patient evaluations of an identical health state that is characterized by a specific illness. These approaches include e.g. that patients adapt to their deteriorated health states, that different assumptions about the recency of the health state are made, that people forget to consider obvious aspects of unfamiliar health states and many others (Brazier, 2008; Dolan & Kahneman, 2008; Ubel *et al.*, 2001, 2003). Most of these explanations explicitly or implicitly refer to observed discrepancies in utility weights elicited by approaches, usually employing SG and TTO, where often a specific state of ill health is described and evaluated as a whole by patients and non-patients (de Wit *et al.*, 2000; Peeters & Stiggelbout, 2010). In contrast, we quantify the effects of selected underlying characteristics on the overall utility of an ill state of health. More specifically, instead of providing a description of a health state and measuring the utility associated with the description as a whole, we include all given information on a health state, i.e. the attributes and according levels, in our model specifications to identify the sources of potential discrepancy.[6] As a result, we believe that none of the mentioned approaches

[6]Note that one could of course also analyze the discrepancies between patients and non-patients with regard to the valuation of disease symptoms or consequences by employing TTO and SG. The main advantage of using a DCE in this context is the comparatively low cognitive burden that is placed on

Table 4.6: MNL regression results

Category	Coefficient	Model 1	Model 2	Model 3	Model 4
Dummy	JOB_{YES}	-0.122*	0.252***	-0.205***	-0.314*
coeff.	FAT5	-1.630***	-1.733***	-	-
	FAT15	-0.701***	-0.662***	-	-
	FMS5	-0.750***	-0.860***	-	-
	FMS15	-0.902***	-0.763***	-	-
	PN5	-2.659***	-2.699***	-	-
	PN15	-1.452***	-1.778***	-	-
	DUR1	-1.519***	-1.573***	-	-
	DUR3	0.206	0.096	-	-
Quant.	FAT	-	-	0.072***	0.072***
coeff.	FMS	-	-	0.032***	0.032***
	PN	-	-	0.116***	0.116***
	DUR	-	-	0.404***	0.405***
Patient	$JOB \times PAT$	-	-	0.286***	0.277***
int.	$FAT \times PAT$	-	-	-0.005	-0.005
	$FMS \times PAT$	-	-	0.001	0.001
	$PN \times PAT$	-	-	-0.002	-0.002
	$DUR \times PAT$	-	-	-0.031	-0.031
Job	$JOB \times INC$	-	-	-	7.12E-05***
int.	$JOB \times SEX$	-	-	-	0.035
	$JOB \times AGE$	-	-	-	0.002
	$JOB \times HH$	-	-	-	-0.099*
	$JOB \times EMP$	-	-	-	0.012
	N	200	227	427	427
	n	6400	7262	13662	13662
	χ^2	900.07***	1042.40***	1821.20***	1844.49***
	Pseudo-R^2	0.380	0.377	0.367	0.368

* Significant at 10%, ** 5%, *** 1%. Model 1: Dummy coded and non-patients. Model 2: Dummy coded and patients. Model 3: Quantitative coefficients, patient interactions and full sample. Model 4: Quantitative coefficients, patient interactions, job interactions with respondent characteristics and full sample.

in this strand of literature can adequately explain why we find a significant discrepancy between patients and non-patients with regard to the valuation of being able to work. Although it is generally believed that unemployment has a negative effect on life satisfaction (Kassenboehmer & Haisken-DeNew, 2009; Winkelmann & Winkelmann, 1998), there are no studies, to our knowledge, that investigate the effects of a temporary incapacitation for work on well-being. On the other hand, there are studies that suggest that vocational status is of relevance to patients (Arns & Linney, 1993; Krokavcova *et al.*, 2012). As a result of patient comments and in accordance with Fifield *et al.* (1991) we are led to believe that non-patients tend to underrate the importance of being able to work because they are, at least to some extent, oblivious of the social implications of being unemployed, even if the period of unemployment lasts for only three months.

Furthermore, our results suggest that household income is positively correlated with the perceived importance of being able to work. In accordance with economic theory, one might think that this finding points to significant opportunity cost that are related to earnings loss. However, a temporary incapacitation for work is usually paid for by the employer or health insurance in Germany. We offer a different possible explanation. We presume that household income is positively correlated with job responsibility (Fox, 2009) and that people with high-responsibility jobs are less willing or able to 'let go' of their vocational obligations. Our finding that household size is negatively correlated with the perceived importance of being able to work could be explained by the intuition that in larger households, it is likely that another household member can assume the responsibility for earnings. As a consequence, the according respondent is less dependent on his or her own ability to work.

So far, the effects of being incapacitated for work on well-being have been largely neglected in the literature on RA. We present evidence that while typical symptoms of RA are associated with a similar utility value by both patients and non-patients, non-patients tend to undervalue the ability to work in comparison. These findings should be considered when RA health states are evaluated, especially when non-patient proxies are used.

the respondents, which is an important feature in the realm of patient related research (Bansback *et al.*, 2012).

5 A comparison of discrete choice and best-worst scaling

5.1 Introduction

Discrete choice experiments (DCEs) have received considerable interest in health economics over the past two decades and have been used to elicit preferences from various groups of respondents in many different contexts to inform policy decisions (de Bekker-Grob *et al.*, 2012). In particular, their theoretical foundation on the well tested and documented random utility theory (McFadden, 1974) makes DCEs more appealing than other stated preference techniques, e.g. the seemingly similar conjoint analysis (Louviere *et al.*, 2010). It is probably due to the increasing popularity of DCEs that also a large number of studies have been published over the past twenty years, which aim at advancing specific aspects of DCEs. For instance, Johnson *et al.* (2013) provide a comprehensive overview of a strand of literature that aims at improving the underlying experimental designs of DCEs. In addition, progress has been made with regard to the applied models to analyze DCE data. Whereas the multinomial logit (MNL) model was once considered to be the workhorse for choice data analysis, alternative models that relax the MNL's assumptions of the independence of irrelevant alternatives (IIA) and the absence of taste as well as scale heterogeneity across respondents have been developed and advocated (Keane & Wasi, 2013).

Despite the remarkable advancements in the realm of DCEs, there are shortcomings that are rooted in the method itself and probably cannot be overcome within the conventional framework. Particularly, Flynn *et al.* (2007) criticize two major points. First, the estimated parameters in a traditional DCE do not allow any substantiated inference, without further assumptions, about the impact of the selected attributes themselves because one must specify a reference level in each attribute domain. Thus, meaningful interpretations of estimated DCE coefficients for dummy coded levels are limited to intra attribute comparisons. Second, DCEs are criticized for not being efficient with regard to the amount of preference information that is elicited per choice task as respondents are asked to just pick one alternative. In order to mitigate these issues, the authors advocate using the relatively novel method best-worst scaling (BWS) to investigate preferences in health economic and other contexts.

'A comparison of discrete choice and best-worst scaling' is joint work with Stefan Felder and Malte Wolff.

BWS was developed by Louviere & Woodworth (1990) and first applied by Finn & Louviere (1992). Flynn (2010) distinguishes between three different choice tasks that are subsumed under the term BWS and he refers to them as BWS cases 1, 2 and 3. The statistical properties of cases 1 and 2, the so-called object and profile cases, were proven by Marley & Louviere (2005) and Marley *et al.* (2008), respectively. Since then, interest in these preference elicitation techniques has risen and a number of applications were published (Coast *et al.*, 2008; Erdem & Rigby, 2013; Ratcliffe *et al.*, 2011). Although the mathematical properties of BWS case 3 have not yet been proven, Lancsar *et al.* (2013) propose different ways of analyzing so-called best worst discrete choice experiment data and provide promising results of one of the first applications of this method.

Due to the fact that DCEs share many similarities with BWS, especially with the BWS cases 2 and 3, researchers have recently become increasingly interested in investigating potential differences between the two methods with regards to empirical results. To the authors' knowledge, Potoglou *et al.* (2011) were the first to compare DCE and BWS results. They found that the normalized DCE and BWS coefficients from their models explaining social care related decisions were not significantly different. In contrast to this result, Whitty *et al.* (2013) as well as Severin *et al.* (2013) found significant differences between (normalized) DCE and BWS coefficients in health economic contexts. Flynn *et al.* (2011) added to the inconclusiveness regarding the comparability of DCE and BWS results. They investigated preference weights in a quality of life setting and reported that while 95 (30% of the sample) appeared to use equal weighting on attributes when answering the DCE and the BWS, 71 respondents (23% of the sample) appeared not to use equal weighting.

It is not yet well understood how and why BWS and DCEs can yield remarkably different empirical results in similar research contexts. In many cases either method can be used to investigate the same research question, so applied researchers are often faced with the conundrum to decide between the two methods. The aim of this paper is to shed more light on this issue. Therefore, we used a similar approach as Whitty *et al.* (2013): We estimated MNL, scaled multinomial logit (SMNL), mixed logit (MIXL) as well as generalized multinomial logit (GMNL) models for a DCE and a BWS profile task (case 2). Subsequently, we compared the results to identify substantial differences. In particular, we used the data of a study that featured a DCE in order to investigate the preferences of the German public with regards to consequences of rheumatoid arthritis (RA)[1] and conducted a one profile BWS online survey with the same underlying orthogonal main effects plan (OMEP) as the DCE.

5.2 DCE and BWS task

Both the DCE and the BWS task included the same five attributes and their levels (one two-level attribute and four three-level attributes). Five expert discussion sessions and a

[1] See chapter 4 for a more thorough discussion of the DCE data.

literature review constituted the basis for the selection of the attributes. Table 5.1 shows all attributes and levels.

Table 5.1: Attributes and levels in DCE and BWS

Attribute	Levels
Inability to work for three months (JOB)	yes (YES) no (NO)
Severe fatigue due to RA (FAT)	all days per month (FAT30) about 15 days per month (FAT15) about 5 days per month (FAT5)
Severe problems with buttoning a shirt or blouse (FMS)	all days per month (FMS30) about 15 days per month (FMS15) about 5 days per month (FMS5)
Severe pain due to RA (PN)	all days per month (PN30) about 15 days per month (PN15) about 5 days per month (PN5)
Duration of treatment until condition starts improving (DUR)	1 month (DUR1) 3 months (DUR3) 6 months (DUR6)

We generated one OMEP with 16 profiles for both the DCE and the BWS task. For the DCE, the 16 OMEP-profiles were first randomly paired with 16 copies of themselves, ensuring that the same profiles were not paired, and then manually re-paired to avoid implausible choice sets. This led to a non-d-efficient binary design; however, we were satisfied with its response efficiency (Johnson *et al.*, 2013). In addition, two rationality tests with one dominant alternative were included. To make the choice task more intuitive for the DCE respondents, they were asked to choose the, in their opinion, inferior alternative in each binary choice set instead of the superior one. For the BWS task, we adopted the 16 OMEP-profiles as they were and asked the respondents to make two choices in each profile; they were asked to choose the attribute-level combination that they perceived to be the 'the least evil' and the one that they perceived to be 'the greatest evil'. In order to avoid semantic misunderstandings, we will from now on use the term 'description' for attribute-level combination. Accordingly, every BWS profile comprises five descriptions. Tables 5.2 and 5.3 show all DCE and BWS profiles.

Table 5.2: DCE choice sets

Choice	State A					State B				
1	NO	FAT30	FMS15	PN30	DUR1	YES	FAT5	FMS30	PN30	DUR6
2	Rationality test[a]									
3	NO	FAT15	FMS5	PN30	DUR1	YES	FAT30	FMS30	PN15	DUR1
4	NO	FAT15	FMS30	PN5	DUR3	YES	FAT5	FMS5	PN30	DUR3
5	YES	FAT5	FMS5	PN5	DUR1	NO	FAT30	FMS5	PN5	DUR6
6	YES	FAT15	FMS15	PN5	DUR6	NO	FAT5	FMS5	PN5	DUR1
7	NO	FAT5	FMS30	PN5	DUR1	NO	FAT5	FMS5	PN15	DUR6
8	NO	FAT5	FMS5	PN5	DUR1	YES	FAT30	FMS5	PN5	DUR3
9	YES	FAT5	FMS5	PN30	DUR3	NO	FAT5	FMS15	PN15	DUR3
10	YES	FAT5	FMS15	PN5	DUR1	YES	FAT15	FMS5	PN15	DUR1
11	NO	FAT30	FMS5	PN5	DUR6	NO	FAT30	FMS15	PN30	DUR1
12	YES	FAT5	FMS30	PN30	DUR6	YES	FAT15	FMS15	PN5	DUR6
13	YES	FAT30	FMS30	PN15	DUR1	NO	FAT15	FMS5	PN30	DUR1
14	NO	FAT5	FMS5	PN15	DUR6	YES	FAT5	FMS15	PN5	DUR1
15	YES	FAT15	FMS5	PN15	DUR1	NO	FAT5	FMS30	PN5	DUR1
16	YES	FAT30	FMS5	PN5	DUR3	NO	FAT15	FMS30	PN5	DUR3
17	NO	FAT5	FMS15	PN15	DUR3	YES	FAT5	FMS5	PN5	DUR1
18	Rationality test[a]									

[a] One alternative is dominant.

Table 5.3: BWS profiles

1	NO	FAT30	FMS15	PN30	DUR1
2	NO	FAT15	FMS5	PN30	DUR1
3	NO	FAT15	FMS30	PN5	DUR3
4	YES	FAT5	FMS5	PN5	DUR1
5	YES	FAT15	FMS15	PN5	DUR6
6	NO	FAT5	FMS30	PN5	DUR1
7	NO	FAT5	FMS5	PN5	DUR1
8	YES	FAT5	FMS5	PN30	DUR3
9	YES	FAT5	FMS15	PN5	DUR1
10	NO	FAT30	FMS5	PN5	DUR6
11	YES	FAT5	FMS30	PN30	DUR6
12	YES	FAT30	FMS30	PN15	DUR1
13	NO	FAT5	FMS5	PN15	DUR6
14	YES	FAT15	FMS5	PN15	DUR1
15	YES	FAT30	FMS5	PN5	DUR3
16	NO	FAT5	FMS15	PN15	DUR3

We developed two very similar online questionnaires for the DCE and the BWS that included the according choice tasks. Both questionnaires included a brief introductory page, a page with a description of RA and the according decision context, the DCE or BWS choices and, finally, conventional socio-economic questions. Figures 5.1 and 5.2 present sample choices for the DCE and BWS task.

A market research company was contracted to recruit 200 respondents for each the DCE and the BWS from their (non-patient) German online panel using quota sampling according to the characteristics sex, age, federal state of residency and level of school education to mimic the German population. Respondents who failed to choose the dominant alternative in both DCE test choice sets were excluded from further analysis. Table 5.4 shows selected descriptive characteristics of the final samples.

5.3 Data analysis

5.3.1 Comparability of BWS and DCE data

In contrast to the DCE, the BWS respondents made two choices in every task. Every respondent n chose the best description (the least evil) and the worst one (the greatest evil) in every profile t. However, it is possible to model the BWS task as a discrete choice (Flynn, 2010; Potoglou *et al.*, 2011; Severin *et al.*, 2013; Whitty *et al.*, 2013). To see how, note that every BWS profile featured five descriptions and the respondents were

Table 5.4: Descriptive statistics of the BWS and DCE sample

	BWS ($N = 199$)	DCE ($N = 200$)
Sex		
Female (%)	50.75%	50.50%
Age		
Mean (yrs)	46.51	49.29
Std. dev.	14.97	14.92
Household income per month		
EUR 0 - <1000	8.54%	5.50%
EUR 1000 - <2000	25.63%	16.50%
EUR 2000 - <3000	17.59%	22.00%
EUR 3000 - <4000	13.57%	13.00%
EUR 4000 - <5000	6.53%	9.00%
>= EUR 5000	6.53%	10.00%
N/A	21.61%	24.00%
Household size		
1 person	18.59%	13.50%
2 persons	41.71%	44.00%
3 or more persons	39.70%	42.50%
School education		
No certificate	0.50%	0.00%
Lower sec. educ.	30.65%	24.50%
Middle school	30.66%	34.50%
A level	34.68%	36.50%
Other	1.51%	2.00%
N/A	2.01%	2.50%
Occupation		
Employed	54.77%	44.50%
Unemployed	7.04%	7.00%
Apprentice	2.01%	0.50%
Civil servant	4.52%	7.00%
Retired	18.59%	23.50%
Self-employed	4.02%	8.00%
Pupil/Student	4.53%	5.00%
Other	3.52%	2.00%
N/A	1.01%	2.50%

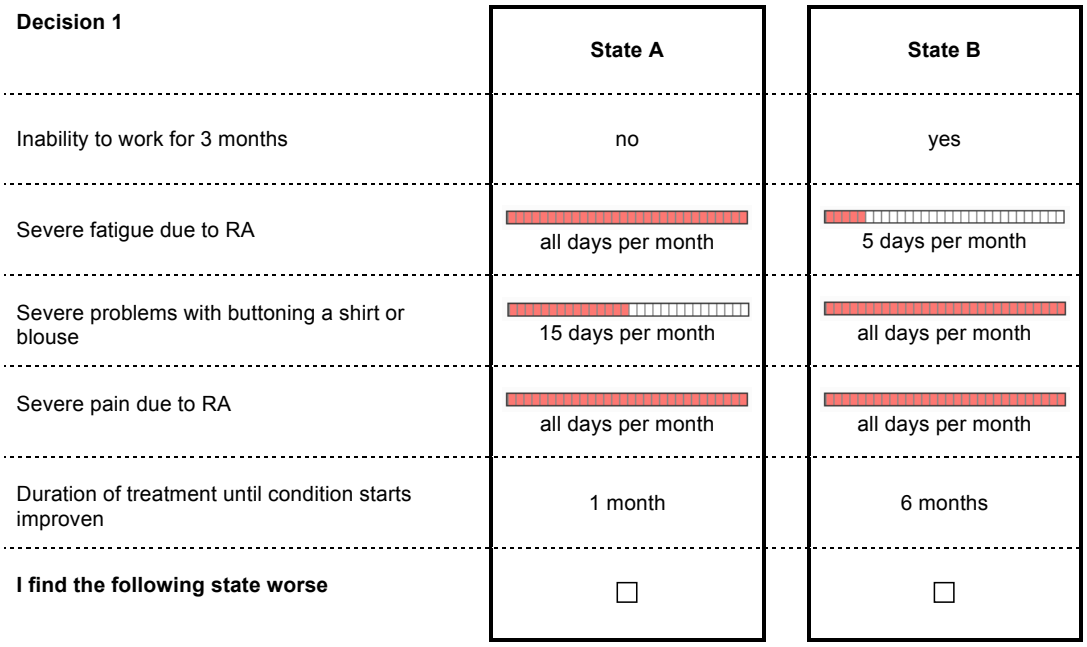

Figure 5.1: Example of a DCE choice set (original in German)

not allowed to choose the identical description as the best and the worst one. Thus, the respondents had to make a discrete choice among $5 \times 4 = 20$ best-worst combinations in any given profile. In particular, the assumed utility function of respondent n for the best-worst combination ij, where i is the best description and j the worst one, in any given profile t can be written as follows:

$$U_{ijnt} = [\beta_i \times D] - [\beta_j \times D] + \epsilon_{ijnt}, \tag{5.1}$$

where D is a dummy variable that takes the value of 1 if the description is equivalent to either i or j in the according best-worst combination and 0 otherwise. Note that every i and j corresponds to one level of an attribute that is given in any profile t. For instance, i or j may correspond to 'severe fatigue on all days per month' ($FAT30$) or 'severe pain on all days per month' ($PN30$) in BWS profile 1, which is depicted in figure 5.2. To illustrate, we provide an excerpt of the resulting structure of the BWS data in table 5.5. In particular, table 5.5 shows the 20 possible best-worst combinations for BWS profile 1. It can be seen in the column labeled 'altchosen' that the respondent with the 'id' 2395 chose the best-worst combination in the first row. More specifically, the values of 1 and -1 in columns JOB_{no} and $FAT30$ in the first row mean that this respondent chose 'not on sick leave' as the best description and 'all days per month: severe fatigue due to rheumatoid arthritis' as the worst one.

Although we can model the BWS task as a discrete choice this way, we do not neglect the additional information that it provides. While in conventional DCEs the intra attribute dummy variables are linearly dependent, this is not the case in the BWS data. That is

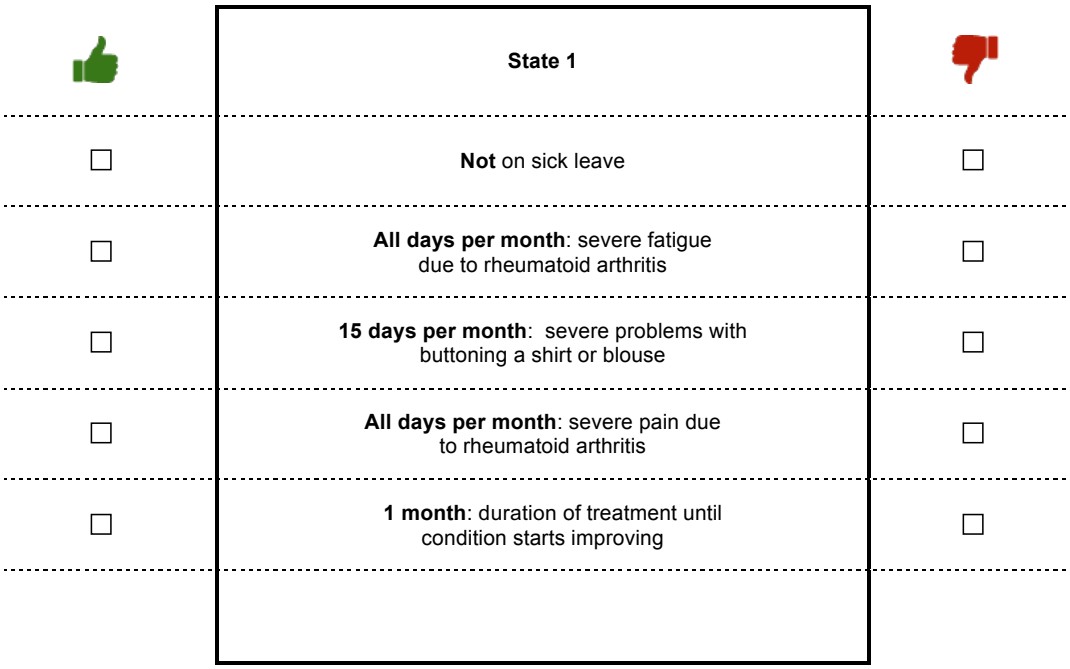

Figure 5.2: Example of a BWS profile (original in German)

because of the assumed data generating process in equation 5.1. The respondents are asked to compare levels of different attributes with each other as opposed to DCEs in which utility differences between levels within an attribute are elicited, but not between levels across different attributes (Flynn *et al.*, 2007). Thus, one needs to specify only one reference level (of any attribute) in BWS models as opposed to the need for specifying one reference level for every attribute in DCE models. As a consequence, the BWS models include more regressors than the DCE models. To mitigate this comparability problem, we follow the approach of Potoglou *et al.* (2011) and subtract, for each attribute, the estimated coefficient for the lowest level from all the higher ones to obtain utility difference estimates between intra attribute levels that are comparable to DCE coefficients.

Furthermore, it is noteworthy that equation 5.1 depicts the assumed data generating process for the MNL model. In the more sophisticated models we use variations of this process that account for scale and/or taste heterogeneity across respondents, which we will discuss in more detail in the following subsection.

Table 5.5: Structure of the BWS data

id	choset	choid	JOB_{yes}	JOB_{no}	FAT5	FAT15	FAT30	FMS5	...	DUR6	altchosen
2395	1	1	0	1	0	0	-1	0	...	0	1
2395	1	1	0	1	0	0	0	0	...	0	0
2395	1	1	0	1	0	0	0	0	...	0	0
2395	1	1	0	1	0	0	0	0	...	0	0
2395	1	1	0	-1	0	0	1	0	...	0	0
2395	1	1	0	0	0	0	1	0	...	0	0
2395	1	1	0	0	0	0	1	0	...	0	0
2395	1	1	0	0	0	0	1	0	...	0	0
2395	1	1	0	-1	0	0	0	0	...	0	0
2395	1	1	0	0	0	0	-1	0	...	0	0
2395	1	1	0	0	0	0	0	0	...	0	0
2395	1	1	0	0	0	0	0	0	...	0	0
2395	1	1	0	-1	0	0	0	0	...	0	0
2395	1	1	0	0	0	0	-1	0	...	0	0
2395	1	1	0	0	0	0	0	0	...	0	0
2395	1	1	0	0	0	0	0	0	...	0	0
2395	1	1	0	-1	0	0	0	0	...	0	0
2395	1	1	0	0	0	0	-1	0	...	0	0
2395	1	1	0	0	0	0	0	0	...	0	0
2395	1	1	0	0	0	0	0	0	...	0	0

This table shows the data of BWS profile 1 ('choset') for the respondent with the 'id' 2395. This profile is depicted in figure 5.2. A value of 1 in the column 'altchosen' shows which of the 20 best-worst combination the respondent chose. The levels FMS15, FMS30, PN5, PN15, PN30, DUR1 and DUR3 are omitted due to space restrictions. See table 5.1 for the descriptions of the attributes JOB, FAT, FMS, PN and DUR and their according levels. The shown data format is required for all relevant models in the statistical package STATA. Other programs may require a different format.

5.3.2 Applied models

In order to be able to evaluate the differences between DCE and BWS results more thoroughly, we estimated four different models for the two preference elicitation methods. First, we estimated the traditional workhorse of choice analysis, the MNL model. In the MNL model the nth individual associates utility U with choosing the alternative i in the choice scenario t:

$$U_{int} = \beta x_{int} + \epsilon_{int}, \tag{5.2}$$

where β is a $1 \times k$ weight vector and x is a $k \times 1$ vector with $k = 1, ..., K$ explanatory variables that may e.g. include alternative specific constants, attributes, dummy or effects coded levels of attributes or interaction terms between attributes and socio-economic variables of the respondents. The MNL's major advantage is that it provides a closed form expression for choice probabilities because the error terms ϵ_{int} are assumed to be i.i.d. extreme value type I (McFadden, 1974):

$$P_{int} = \frac{exp(U_{int})}{\sum_{j=1}^{J} exp(U_{jnt})} \qquad \forall j \in j = 1, ..., J; i \neq j. \tag{5.3}$$

However, it suffers from the fact that it imposes strong assumptions on the choice behavior of the respondents. Particularly, the MNL has been criticized for the assumption of a very specific distribution of the error terms and because the ratio of two choice alternatives is only dependent on their attributes and thus remains constant when new alternatives are added, the so-called IIA property (Keane, 1997). Furthermore, the MNL's underlying utility structure as shown in equation 5.2 does not allow for preference heterogeneity of observed variables, which is illustrated by the β-vector's independence of the nth individual.[2] We used the traditional MNL as a starting point for our analysis of differences between BWS and DCE. We estimated MNL models for both methods and let all attribute levels enter the utility specification in equation 2 additive-linearly and dummy coded. For the DCE we chose the maximum level in every attribute domain to be the reference case[3] and for the BWS we chose PN30 to be the reference case. We did not include alternative specific constants in the DCE model because it was generic and did not feature meaningful labels to distinguish the two shown alternatives in the choice sets.

As a next step, we estimated the MIXL model for both the DCE and the BWS task, as proposed by Revelt & Train (1998), to allow for preference heterogeneity in observable

[2]The MNL does, however, allow for heterogeneity of unobserved variables across respondents in the error terms.

[3]For instance, for the attribute 'pain' (PN) we chose PN30 to be the reference case to which all other levels in this domain are compared to.

variables and to overcome the limiting IIA assumption. In the MIXL model, person nth's utility U of alternative i in choice scenario t is given by

$$U_{int} = (\beta + \eta_n)x_{int} + \epsilon_{int}. \tag{5.4}$$

In this model, respondent specific deviations from the utility mean β are introduced with the vector η_n. While ϵ_{int} is still assumed to be i.i.d. extreme value type I, η_n is assumed to be multivariate normal, $MVN(0, \sum)$, with a diagonal variance matrix $\sum$. Through the introduction of η_n, the MIXL allows for correlation in tastes across alternatives and thus avoids the IIA assumption. More specifically, $(\beta + \eta_n) = \beta_n$ is unobserved and assumed to vary with the density function $f(\beta_n|\theta^*)$, where θ^* represents the true parameters underlying the distribution. In our application, it is assumed that θ^* contains the means and standard deviations[4] of β_n and is estimated to explain the probability that alternative i is chosen accordingly:

$$P(i|\theta)_{int} = \int \frac{exp(U_{int})}{\sum_{j=1}^{J} exp(U_{jnt})} f(\beta_n|\theta)d\beta_n \qquad \forall j \in j = 1, ..., J; i \neq j. \tag{5.5}$$

In practice, most applications use an arbitrary number of D Halton draws ${\eta^d}_{d=1,...,D}$ from the assumed multivariate normal distribution to obtain simulated probabilities:

$$P(i|x)_{int} = \frac{1}{D} \sum_{d=1}^{D} \frac{exp((\beta + \eta^d)x_{int})}{\sum_{j=1}^{J} exp((\beta + \eta^d)x_{jnt})} \qquad \forall j \in j = 1, ..., J; i \neq j. \tag{5.6}$$

Analogously to the MNL specifications, we included all attribute levels additively and linearly in equation 5.4 with the same reference cases. As we expected that our respondents might exhibit preference heterogeneity with regard to all attributes, we specified all regressors to be random.[5] We used the STATA module developed by Hole (2007) with 250 Halton draws to estimate the MIXL models.

The analysis of the differences between the DCE and the BWS task with regard to the exhibited level of preference heterogeneity across respondents is one of this study's major contributions to the existing literature. The fact that both the DCE and the BWS task had the same underlying OMEP, featured a very similar online questionnaire and were conducted with participants of the same panel with similar socio-economic characteristics

[4] Note that we assume $\sum$ to be diagonal so that we neglect any possible covariances.

[5] It is possible to include random and non-random variables in the same MIXL specification, where the latter are assumed to be unconditional on θ and thus resemble traditional MNL coefficients. For instance, we include both random and non-random variables in chapter 3.

leads us to the proposition that any differences between the resulting degrees of preference heterogeneity are induced by the chosen method.

However, the MIXL was criticized for being misspecified because a major part of preference heterogeneity could be attributed to differently scaled error terms across respondents (Louviere & Meyer, 2007; Louviere *et al.*, 1999, 2002, 2008). Thus, it is argued that for some respondents choices are more random than for others. This would imply that the MIXL does not identify real taste heterogeneity, i.e. different utility evaluations of attribute levels across respondents, but a confounded measure of scale and taste heterogeneity.

In order to address this issue and to be able to dissect any potential preference heterogeneity differences between the DCE and BWS task into taste and scale heterogeneity dissimilarities, we also estimated SMNL and GMNL models, as proposed by Fiebig *et al.* (2010). In the former scale heterogeneity is modeled as the scaling factor σ_n, which is a random scalar that is assumed to be distributed log-normal with standard deviation τ and mean 1. The latter constraint is necessary for identification. The SMNL can be written as follows:

$$U_{int} = (\sigma_n \beta) x_{int} + \epsilon_{int}. \tag{5.7}$$

Finally, we estimated the GNML that nests the SMNL and the MIXL models and was developed by Fiebig *et al.* (2010):

$$U_{int} = [\sigma_n \beta + \gamma \eta_n + (1-\gamma)\sigma_n \eta_n] x_{int} + \epsilon_{int}, \tag{5.8}$$

where η_n and σ_n are the relevant taste and scale heterogeneity parameters from the MIXL and SMNL models. γ describes how the two sources of heterogeneity are related to each other. While Fiebig *et al.* (2010) constrained γ to be between 0 and 1, we followed the advice of Keane & Wasi (2013) and did not do so in our estimations. Note that by setting the variance-covariance matrix of η_n to 0, one obtains the SMNL model and by setting the scale parameter σ_n to 1, one obtains the MIXL model. In particular, we used the STATA module developed by Gu *et al.* (2013) with 250 Halton draws to estimate the SMNL and GMNL models, assuming that all parameters are random in the latter:

$$P(i|x)_{int} = \frac{1}{D} \sum_{d=1}^{D} \prod_{t} \prod_{i} \frac{exp([\sigma^d \beta + \gamma \eta^d + (1-\gamma)\sigma^d \eta^d] x_{int})}{\sum_{j=1}^{J} exp([\sigma^d \beta + \gamma \eta^d + (1-\gamma)\sigma^d \eta^d] x_{jnt})}, \tag{5.9}$$

$\forall j \in j = 1, ..., J; i \neq j$, where the scale parameter σ^d is simulated by

$$\sigma^d = exp(\bar{\sigma} + \tau v^d). \tag{5.10}$$

Here, $\bar{\sigma}$ is a constant and is set to $-ln(\frac{1}{N}\sum_{i=1}^{N} exp(\tau v_i^d))$. This corresponds to setting the mean of σ_n equal to 1 in the simulated data and its standard deviation to τ (Fiebig *et al.*, 2010). In the applied STATA module, a combination of Halton and pseudo-random draws are used to generate the $N(0,1)$ scalar v^d (Gu *et al.*, 2013). These restrictions on the scale parameter are necessary for the identification of β. As in the MIXL models, all attribute-levels enter the GMNL models as random.

5.4 Results

While in both samples the respondents found it harder to take decisions than to understand the according choice task, the BWS respondents rated their task to be more difficult than the DCE respondents: 25.13% of the BWS respondents found their task hard or very hard whereas only 11.00% of the DCE respondents rated their task as hard or very hard. This finding is in line with the results reported by Severin *et al.* (2013) and Whitty *et al.* (2013) and provides tentative evidence against the claim of Flynn *et al.* (2007) that the BWS task is potentially easier than the DCE.

Table 5.6: Perceived difficulties of BWS and DCE

	BWS sample	DCE sample
Difficulty comprehending task		
Not hard at all	23.12%	31.50%
Not very hard	51.76%	57.50 %
Hard	22.11%	11.00%
Very hard	3.02%	0.00%
Difficulty taking decisions		
Not hard at all	7.04%	7.50%
Not very hard	40.20%	59.00%
Hard	46.23%	30.50 %
Very hard	6.53%	3.00%

N(BWS) = 199. N(DCE) = 200.

The results of all estimated models are presented in tables 5.7 and 5.8. While table 5.7 comprises all estimated β-coefficients and statistics regarding model fit, table 5.8 presents all estimated standard deviations of the coefficients in the MIXL and GMNL models and the relevant (scale) heterogeneity related parameters τ and γ, if applicable.

Regarding the DCE coefficients presented in table 5.7, there are three notable deviations from what one would intuitively expect. First, JOB_{yes} is negative in all models and significantly different from zero in three of them, which means that on average, the (non-patient) respondents associated a higher level of utility with being on sick leave for three months than not being on sick leave. A possible reason for this observation could be that non-patients do not value the ability to work very highly.[6] Second, the coefficient for FMS15 is lower than the one for FMS5. This seems unreasonable since having problems with buttoning a shirt on 15 days per month is clearly inferior to having problems with it on only 5 days per month. However, Wald tests on the equality of the FMS15 and FMS5 coefficient values revealed that the difference between the two is not significantly different from zero in any of the four DCE models. Thus, the respondents in the DCE just did not associate meaningfully different levels of utility with the two levels.[7] Finally, DUR3 is not significant in any model, which can be interpreted similarly to the observation in the FMS attribute domain: The respondents did not distinguish between a period of suffering for three (DUR3) and six months (DUR6) until the condition starts to improve from a utility perspective. The log-likelihood values as well as the Akaike (AIC) and Bayesian Information Criterion (BIC) values suggest that the GMNL model fits best, followed by the MIXL and then the SMNL model, while the traditional MNL model falls behind.

The BWS coefficients given in table 5.7 are all in line with prior expectations. All coefficients are significant and exhibit the expected intra attribute preference ordering, while the highest degree of the attribute pain (PN30) is perceived to be the worst attribute-level combination.[8] Interestingly, the MIXL model performs better than the more sophisticated GMNL model for the BWS with regard to model fit according to the AIC and BIC. However, both models outperform the SMNL and MNL models, where the latter ranks last.

[6] See chapter 4 for a more comprehensive discussion.

[7] Note that the respondents did, however, associate a clearly lower level of utility with FMS30 compared to either FMS5 and FMS15.

[8] This can be seen from the fact that all estimated coefficients are positive and the reference case is PN30.

Table 5.7: Estimated coefficients of MNL, SMNL, MIXL and GMNL models for DCE and BWS

	DCE				**BWS**			
	MNL	SMNL	MIXL	GMNL	MNL	SMNL	MIXL	GMNL
Coefficients								
JOB_{yes}	-0.12*	-0.80***	-0.16	-0.23*	1.68***	2.33***	2.89***	3.12***
JOB_{no}	-	-	-	-	2.45***	3.25***	3.90***	3.75***
FAT5	-1.63***	-3.66***	-2.13***	-2.51***	1.69***	2.20***	2.64***	2.94***
FAT15	-0.70***	-2.03***	-0.87***	-1.15***	0.77***	1.00***	1.07***	1.18***
FAT30	-	-	-	-	0.17***	0.37***	0.23***	0.30***
FMS5	-0.75***	-1.13***	-0.94***	-0.94***	1.79***	2.35***	2.81***	3.15***
FMS15	-0.90***	-1.55***	-1.07***	-1.14***	1.07***	1.53***	1.59***	1.83***
FMS30	-	-	-	-	0.53***	0.91***	0.79***	0.92***
PN5	-2.66***	-4.72***	-3.59***	-3.88***	1.26***	1.70***	1.98***	2.18***
PN15	-1.45***	-2.72***	-1.85***	-2.19***	0.44***	0.59***	0.65***	0.77***
PN30	-	-	-	-	-	-	-	-
DUR1	-1.52***	-3.47***	-2.14***	-2.43***	2.07***	2.77***	2.89***	3.81***
DUR3	0.21	1.16	0.26	0.43	0.88***	1.19***	1.22***	1.30***
DUR6	-	-	-	-	0.20**	0.48***	0.30***	0.57***
Model fit								
LL	-1375.82	-1336.44	-1301.74	-1282.27	-8190.68	-7869.67	-6206.31	-6266.26
N	6400	6400	6400	6400	63680	63360	63360	63360
Parameters	9	10	18	20	13	14	26	28
AIC	2769.64	2692.87	2639.48	2604.54	16406.36	15767.34	12464.62	12588.52
BIC	2830.52	2760.51	2761.23	2739.82	16524.16	15894.14	12700.09	12842.10

* Significant at 10%, ** 5%, *** 1%. The AIC is not adjusted for the number of observations. In the MIXL and GNML models, all regressors are assumed to be random.

Table 5.8: Estimated heterogeneity parameters of MNL, SMNL, MIXL and GMNL models for DCE and BWS

	DCE				**BWS**			
	MNL	SMNL	MIXL	GMNL	MNL	SMNL	MIXL	GMNL
Std. dev.								
JOB_{yes}	-	-	1.01***	0.00	-	-	2.87***	1.02***
JOB_{no}	-	-	-	-	-	-	3.20***	1.17***
FAT5	-	-	-0.60***	0.00	-	-	0.38***	-0.04
FAT15	-	-	-0.05	0.00	-	-	-0.26**	-0.01
FAT30	-	-	-	-	-	-	-0.23*	-0.02
FMS5	-	-	0.87***	0.00	-	-	-0.99***	-0.12***
FMS15	-	-	0.05	0.00	-	-	-0.05	-0.06
FMS30	-	-	-	-	-	-	0.39***	-0.01
PN5	-	-	0.88***	0.00	-	-	0.43***	0.21***
PN15	-	-	0.22	0.00	-	-	0.09	0.26***
PN30	-	-	-	-	-	-	-	-
DUR1	-	-	1.00***	0.00	-	-	2.38***	0.93***
DUR3	-	-	-0.03	0.00	-	-	2.11***	0.90***
DUR6	-	-	-	-	-	-	2.71***	1.08***
Scale param.								
τ	-	0.96***	-	0.46***	-	0.85***	-	0.52***
γ	-	-	-	35270.52	-	-	-	-7.58***

* Significant at 10%, ** 5%, *** 1%. The sign of the estimated standard deviations is irrelevant; they can be interpreted as being positive (Hole, 2007). The parameter γ is not constrained to be between 0 and 1 (Gu *et al.*, 2013; Keane & Wasi, 2013).

In order to compare the estimated DCE coefficients with the BWS ones, we took the approach proposed by Potoglou *et al.* (2011) to rescale the latter. This means that we first subtracted in each BWS attribute domain the worst level (e.g. FAT30) from the best (e.g. FAT 5) to obtain intra attribute utility differences that are comparable to the DCE coefficients. Second, we set one attribute-level combination in the DCE and the BWS to 1 and rescaled all other coefficients accordingly. We chose FAT5 for this purpose. This procedure was necessary because the BWS and DCE have potentially dissimilar underlying utility scales and thus cannot be compared directly (Louviere & Swait, 1993). The rescaled coefficients are depicted in figure 5.3 for all four models. In contrast to the findings of Potoglou *et al.* (2011) and in accordance with Severin *et al.* (2013) and Whitty *et al.* (2013), we find that there are significant differences between all rescaled DCE and BWS coefficients in every model. As noted earlier, there are even different preference orderings in the attribute domains JOB and FMS.

Furthermore, we used scatter plots, including an ordinary least squares analysis, to compare the rescaled BWS and DCE coefficients per model and to investigate whether accounting for scale or taste heterogeneity in the different models mitigates the dissimilarities between the rescaled coefficients. Figure 5.4 provides tentative evidence for the latter conjecture because the R-squared values increased from 0.33 (MNL) to 0.41 (GMNL) when accounting for both scale and taste heterogeneity. The R-squared also increased to 0.34 when we only accounted for scale heterogeneity (SMNL) and to 0.38 in the MIXL with its confounded taste and scale heterogeneity specification.

In table 5.8 we provide more information on the importance of scale and taste heterogeneity for both the DCE and the BWS task. In fact, the table shows that scale heterogeneity is important in both the DCE and the BWS task as for both methods the SMNL and the GMNL models report a significant τ parameter. However, the BWS task seems to be much more prone to taste heterogeneity across respondents than the DCE. This can be inferred from the observation that after controlling for scale heterogeneity in the GMNL models, the standard deviations in the DCE estimation are close to zero while there are eight out of 13 BWS coefficients with significant standard deviations.

Another approach to evaluate the importance of scale and taste heterogeneity in the BWS and DCE tasks is to compare the successive log-likelihood improvements when going sequentially from the MNL to the SMNL as well as to the GMNL model (Fiebig *et al.*, 2010). Table 5.9 provides these log-likelihood percentage improvements for our estimated models as well as for the ones estimated by Whitty *et al.* (2013), though they do not report them explicitly.

According to table 5.9, the total improvement of log-likelihood by accounting for scale and preference heterogeneity in our DCE and BWS task is 7% and 23%, respectively. The percentage improvements in the study of Whitty *et al.* (2013) are considerably smaller with 2% for the DCE and 7% for the BWS. It is, however, remarkable that in both studies the BWS task boasts significantly larger log-likelihood improvements when accounting for scale and taste heterogeneity. Moreover, the inclusion of scale heterogeneity accounts for a much larger fraction of the total log-likelihood improvements in the DCE than in the BWS task: 42% and 44% in the DCE versus 17% and 13% in the BWS task. This

Table 5.9: Log-likelihood improvements when accounting for heterogeneity

	These authors		**Whitty et al.**[a]	
	DCE	BWS	DCE	BWS
MNL LL	-1376	-8190	-3401	-17232
SMNL LL	-1336	-7870	-3373	-17071
GMNL LL	-1282	-6266	-3338	-16022
Δ% MNL to GMNL (1)	-7%	-23%	-2%	-7%
Δ% MNL to SMNL (2)	-3%	-4 %	-1%	-1%
Δ% (2)/(3)[b]	42%	17%	44%	13%

[a] We used the results reported in table 4 in Whitty *et al.* (2013). [b] This indicates the importance of scale heterogeneity because it depicts the fraction of log-likelihood improvement by accounting for scale heterogeneity of the total improvement of log-likelihood by accounting for taste and scale heterogeneity.

supports the previous conjecture that the BWS task seems to be more likely to induce taste heterogeneity. As a matter of fact, it is surprising that by only adding scale heterogeneity the log-likelihood improvements are very similar in the DCE and BWS task: 3% versus 4% in our study and 1% for both methods in Whitty *et al.* (2013).

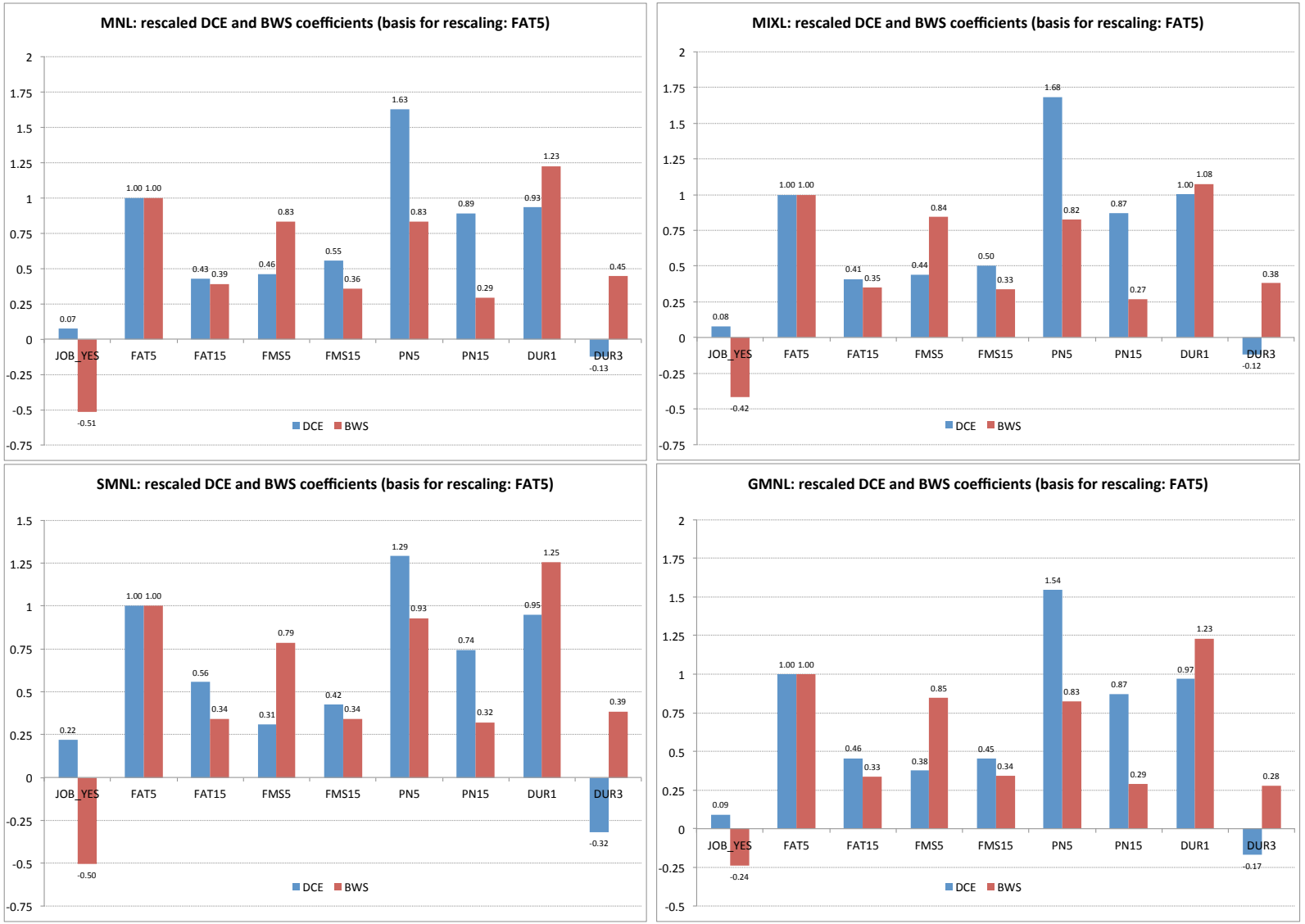

Figure 5.3: Comparison of rescaled DCE and BWS coefficients per estimated model

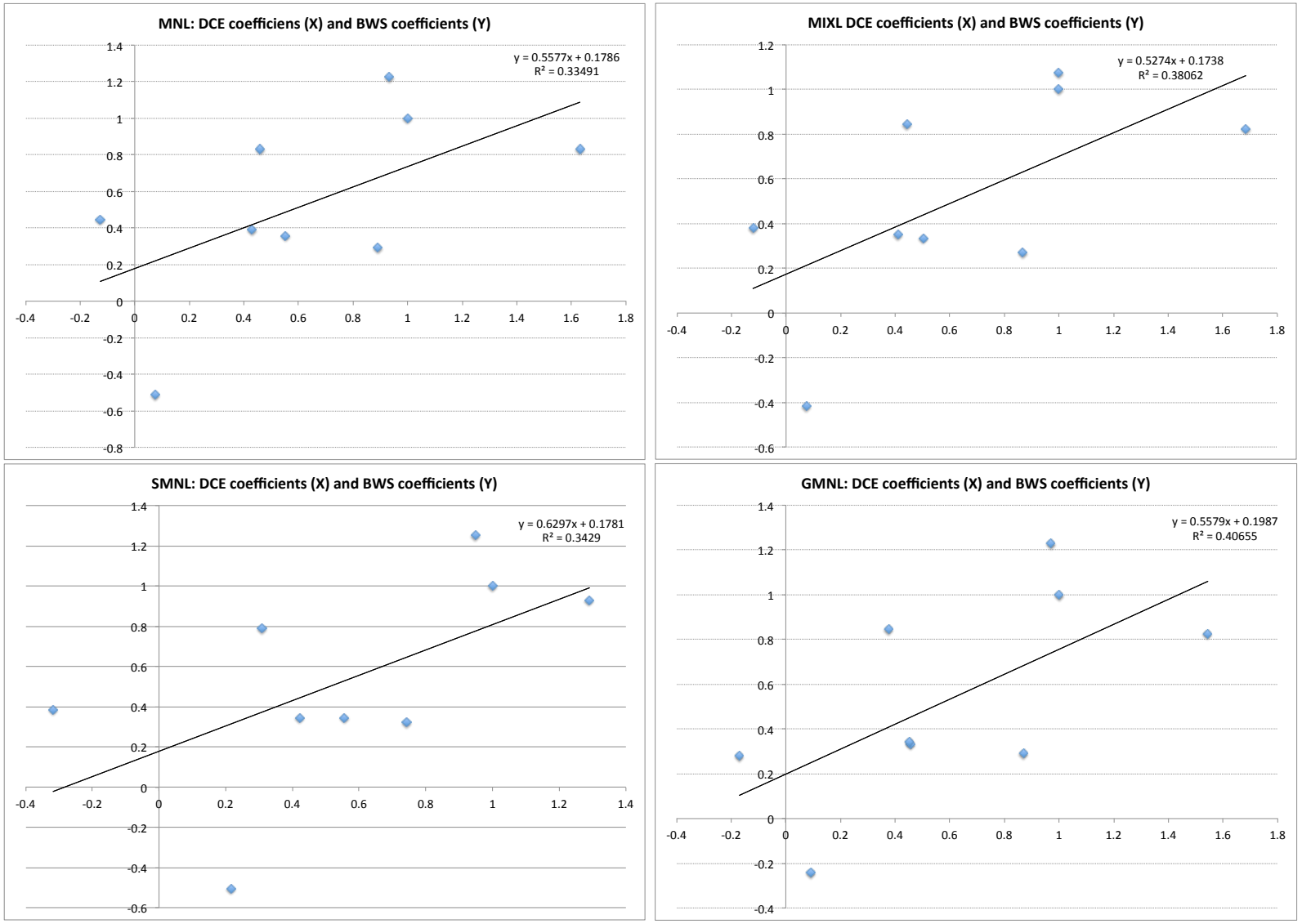

Figure 5.4: Scatter plots for rescaled DCE and BWS coefficients per estimated model including OLS

5.5 Discussion

In this paper we compared the results of a DCE and a BWS task with the same underlying OMEP to shed more light on potential differences between the two methods. We found that the DCE and BWS led to considerably different coefficient estimates and even to different intra attribute preference orderings for two attribute domains. However, while the results of both methods exhibited approximately the same degree of scale heterogeneity, we found tentative evidence that the BWS task induced a larger degree of taste heterogeneity across respondents. Although this result is based on the comparison of only one DCE and one BWS task which comprised about 200 respondents each, the results reported by Whitty *et al.* (2013) follow a very similar pattern as described in the previous section. One might presume that the larger number of alternatives in every choice set in the BWS task is responsible for inducing a larger degree of heterogeneity. However, this conjecture is questionable. For instance, Fiebig *et al.* (2010) note in this context that they had little success trying to explain scale heterogeneity with measures of task complexity that included e.g. the number of attributes and the number of alternatives.

In addition, we found that by accounting for scale and taste heterogeneity, the rescaled DCE and BWS coefficients became more aligned. This indicates that preference heterogeneity is indeed a factor in explaining why DCEs and BWS tasks can produce considerably different results. It is a promising result that the more sophisticated models are capable of mitigating the differences between the two methods which supports the validity of both instruments and encourages further research on models that are even better suited to account for preference heterogeneity.

To conclude, the results of our study do not allow a definite statement about the quality of either method. Yet, we provided useful results for applied researchers who have to decide between conducting a DCE or a BWS task. Taste heterogeneity may be a welcome or unwelcome guest depending on the research question at hand.

6 Conclusion

This thesis comprises four studies in which health related preferences are elicited from the German public. All studies have in common that DCE data is analyzed while they are related to different research questions. The studies presented in chapters 2 and 3 exhibit the highest degree of similarity. Both are aimed at investigating the effect of a specific variable or choice characteristic related with the monetary valuation of a life. While the results of 'Rescuing Schelling's girl' point to a significantly higher level of utility associated with saving an identified over a statistical life, it is found in 'The dead-anyway effect from a societal perspective' that the WTP for a reduction in mortality risk is positively correlated with the level of initial (mortality) risk. Since the estimation procedures in both studies are based on the same data set, the reported WTP values are principally comparable to each other.[1] The reported difference between saving an identified and a statistical life is approximately EUR 2.9 million, whereas it is found that the WTP for a one percentage point decrease in mortality risk increases by EUR 36,000 for every one percentage point increase in initial risk. The derivation of both values depends on the strong assumption of linear part-worth utilities in the range from 0% to 100% initial mortality risk. If we accept this additional assumption, we can argue that the value of a statistical life is $EUR\ 36,000 \times 100 = EUR\ 3.6$ million, which is broadly in line with the findings of Viscusi & Aldy (2003), who report a typical range of USD 4 to 9 million.[2] Thus, a comparison between the results in chapters 2 and 3 suggests that an identified life is valued 1.8 times as much as a statistical life.

The results reported in 'Evaluating the consequences of rheumatoid arthritis' cannot be directly compared to either of the previously discussed studies for two reasons. First, a different DCE was conducted and thus the data sets are not identical, so that scale factor differences prevent direct comparisons (Louviere & Swait, 1993). Second, we refrained from including a monetary attribute in the RA study because the related research question did not make it essential and we could ensure this way that we avoided controversial discussions with patients suffering from RA. Loosely speaking, the studies lack a common denominator to be comparable. However, the result that among all investigated consequences only being unable to work is associated with a different level of utility by patients in comparison to non-patients may be interpreted as evidence for a strong ability of the general public to sympathize with patients regarding states of suffering, even if they are inexperienced. In contrast to this rather convergent choice behavior by the two investigated subgroups, the results provided in chapter 2 are remarkably different: In 'Rescuing

[1] Note in this context that the estimated coefficients for CST are very close to each other in model 2 in chapter 2 and specifications 1 and 2 in chapter 3.

[2] Tsuge *et al.* (2005) use a very similar approach to estimate the value of a statistical life.

Schelling's girl' the estimated coefficients for all attribute levels are more dissimilar in size, and for the CST domain even with respect to significance, across both age groups. In part, this could be explained by the observation that the DCE in chapter 2 was rated more difficult along both dimensions comprehensibility of the task at hand and difficulty to make a decision in the scenarios.[3]

The latter conjecture is supported by the combination of the two results in 'A comparison of discrete choice and best-worst scaling' that the respondents of the BWS task exhibit higher levels of taste heterogeneity and that the BWS is rated to be more difficult along the previously mentioned dimensions.[4] Overall, the results presented in chapter 5 emphasize the need for further research aimed at discovering the reasons for the remarkable differences between BWS and DCE results (see also Severin *et al.*, 2013; Whitty *et al.*, 2013). Although we note in section 5.5 that the more sophisticated models' ability to mitigate these differences is promising evidence for the validity of both experiments, the observed differences between the coefficient estimates of both methods are still too large to be neglected.

To conclude, the intertwinement of microeconomic choice theory and econometric modeling through RUT enables DCEs to produce comparatively convincing results – even in the absence of observable markets. Despite our rather unambiguous results presented in chapters 2 to 4, our research in chapter 5 suggests that a more thorough investigation of the factors that contribute to taste heterogeneity, especially in the GMNL model, is a promising starting point for research in order to eventually shed more light on the underlying differences between DCEs and BWS surveys. Due to the fact that GMNL models do not per se require specifically designed DCEs, available DCE data could be used for these endeavors.

[3] This is inferred from a comparison of tables 2.5 and 4.4.
[4] This can be seen in table 5.6.

Bibliography

Adamowicz, W., Louviere, J. J., & Williams, M. 1994. Combining revealed and stated preference methods for valuing environmental amenities. *Journal of Environmental Economics and Management*, **26**(3), 271–292.

Allenby, G. M., & Rossi, P. 1998. Marketing models of consumer heterogeneity. *Journal of Econometrics*, **89**(1-2), 57–78.

Arns, P. G., & Linney, J. A. 1993. Work, self, and life satisfaction for persons with severe and persistent mental disorders. *Psychiatric Rehabilitation Journal*, **17**(2), 63–79.

Bansback, N., Brazier, J., Tsuchiya, A., & Anis, A. 2012. Using a discrete choice experiment to estimate health state utility values. *Journal of Health Economics*, **31**(2), 306–318.

Bech, M., & Gyrd-Hansen, D. 2005. Effects coding in discrete choice experiments. *Health Economics*, **14**(10), 1079–1083.

Bellavance, F., Dionne, G., & Lebeau, M. 2009. The value of a statistical life: a meta-analysis with a mixed effects regression model. *Journal of Health Economics*, **28**(2), 444–464.

Boyd, N. F., Sutherland, H. J., Heasman, K. Z., Trichtler, D. L., & Cummings, B. J. 1990. Whose utilities for decision analysis? *Medical Decision Making*, **10**(1), 58–67.

Brazier, J. 2008. Valuing health states for use in cost-effectiveness analysis. *Pharmacoeconomics*, **26**(9), 769–779.

Breyer, F., & Felder, S. 2005. Mortality risk and the value of a statistical life: the dead-anyway effect revis(it)ed. *The Geneva Risk and Insurance Review*, **30**(1), 41–55.

Burton, W., Morrison, A., Maclean, R., & Ruderman, E. 2006. Systematic review of studies of productivity loss due to rheumatoid arthritis. *Occupational Medicine*, **56**(1), 18–27.

Campbell, R. C., Batley, M., Hammond, A., Ibrahim, F., Kingsley, G., & Scott, D. L. 2012. The impact of disease activity, pain, disability and treatments on fatigue in established rheumatoid arthritis. *Clinical Rheumatology*, **31**(4), 717–722.

Carmona, L., Cross, M., Williams, B., Lassere, M., & March, L. 2010. Rheumatoid arthritis. *Best Practice & Research Clinical Rheumatology*, **24**(6), 733–745.

Coast, J., Flynn, T. N., Natarajan, L., Sproston, K., Lewis, J., Louviere, J. J., & Peters, T. J. 2008. Valuing the ICECAP capability index for older people. *Social Science & Medicine*, **67**(5), 874–882.

Corso, P. S., Hammitt, J. K., Graham, J. D., Dicker, R. C., & Goldie, S. J. 2002. Assessing preferences for prevention versus treatment using willingness to pay. *Medical Decision Making*, **22**(2), S92–S101.

Covic, T., Adamson, B., & Hough, M. 2000. The impact of passive coping on rheumatoid arthritis pain. *Rheumatology*, **39**(9), 1027–1030.

Davidson, J. D. 1973. Forecasting Traffic on STOL. *Operational Research Quarterly*, **24**(4), 561–569.

de Bekker-Grob, E. W., Ryan, M., & Gerard, K. 2012. Discrete choice experiments in health economics: a review of the literature. *Health Economics*, **21**(2), 145–172.

de Wit, G. A., Busschbach, J. J. V., & de Charro, F. T. 2000. Sensitivity and perspective in the valuation of health status: whose values count? *Health Economics*, **9**(2), 109–126.

Dolan, P., & Kahneman, D. 2008. Interpretations of utility and their implications for the valuation of health. *The Economic Journal*, **118**(525), 215–234.

Elrod, T., & Keane, M. P. 1995. A factor-analytic probit model for representing the market structure in panel data. *Journal of Marketing Research*, **32**(1), 1–16.

Emery, D., & Barron, F. 1979. Axiomatic and numerical conjoint measurement: an evaluation of diagnostic efficacy. *Psychometrika*, **44**(2), 195–210.

Erdem, S., & Rigby, D. 2013. Investigating heterogeneity in the characterization of risks using best worst scaling. *Risk Analysis*, **33**(9), 1728–1748.

Federal Statistical Office. 2009. *Statistical Yearbook 2009*. Federal Statistical Office: Wiesbaden.

Federal Statistical Office. 2010. *Statistical Yearbook 2010*. Federal Statistical Office: Wiesbaden.

Fetherstonhaugh, D., Slovic, P., Johnson, S. M., & Friedrich, J. 1997. Insensitivity to the value of human life: a study of psychophysical numbing. *Journal of Risk and Uncertainty*, **14**, 283–300.

Fiebig, D. G., Keane, M. P., Louviere, J. J., & Wasi, N. 2010. The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. *Marketing Science*, **29**(3), 393–421.

Fifield, J., Reisine, S. T., & Grady, K. 1991. Work disability and the experience of pain and depression in rheumatoid arthritis. *Social Science & Medicine*, **33**(5), 579–585.

Finn, A., & Louviere, J. J. 1992. Determining the appropriate response to evidence of public concern: the case of food safety. *Journal of Public Policy & Marketing*, **11**(2), 12–25.

Flynn, T. N. 2010. Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. *Expert Review of Pharmacoeconomics & Outcome Research*, **10**(3), 259–267.

Flynn, T. N., Louviere, J. J., Peters, T. J., & Coast, J. 2007. Best-worst scaling: what it can do for health care research and how to do it. *Journal of Health Economics*, **26**(1), 171–189.

Flynn, T. N., Peters, T. J., & Coast, J. 2011. Quantifying response shift or adaptation effects in quality of life by synthesising best-worst scaling and discrete choice data. *Journal of Choice Modelling*, **6**, 34–43.

Fox, J. T. 2009. Firm-size wage gaps, job responsibility and hierarchical matching. *Journal of Labor Economics*, **27**(1), 83–126.

Gold, M. R., Siegel, J., Russell, L. B., & Weinstein, M. 1999. *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Green, C., & Gerard, K. 2009. Exploring the social value of health-care interventions: a stated preference discrete choice experiment. *Health Economics*, **18**(8), 951–976.

Grisolía, J. M., & Willisa, K. G. 2011. An evening at the theatre: using choice experiments to model preferences for theatres and theatrical productions. *Applied Economics*, **43**(27), 3987–3998.

Gu, Y., Hole, A. R., & Knox, S. 2013. Fitting the generalized multinomial logit model in Stata. *The Stata Journal*, **13**(2), 382–397.

Hammitt, J. K., & Treich, N. 2007. Statistical vs. identified lives in benefit-cost analysis. *Journal of Risk and Uncertainty*, **35**(1), 45–66.

Hays, R. D., Vickrey, B. G., Hermann, B. P., Perrine, K., Cramer, J., Meador, K., Spritzer, K., & Devinsky, O. 1995. Agreement between self reports and proxy reports of quality of life in epilepsy patients. *Quality of Life Research*, **4**(2), 159–168.

Health Economics. 2013. *Author guidelines*. `http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1099-1050/homepage/ForAuthors.html`. [Accessed: 5 December 2013].

Hensher, D. A. 1994. Stated preference analysis of travel choices: the state of practice. *Transportation*, **21**(2), 107–133.

Hensher, D. A., Rose, J. M., & Greene, W. H. 2007. *Applied choice analysis*. Cambridge University Press: Cambridge.

Hole, A. R. 2007. Fitting mixed logit models by using maximum simulated likelihood. *The Stata Journal*, **7**(3), 388–401.

Hoyos, D. 2010. The state of the art of environmental valuation with discrete choice experiments. *Ecological Economics*, **69**(8), 1595–1603.

Jenni, K. E., & Loewenstein, G. 1997. Explaining the identifiable victim effect. *The Journal of Risk and Uncertainty*, **14**(3), 235–257.

Johnson, F. R., Lancsar, E., Marshall, D., Kilambi, V., Mühlbacher, A., Regier, D. A., Bresnahan, B., Kannien, B., & Bridges, J. F. P. 2013. Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task force. *Value in Health*, **16**(1), 3–13.

Juerges, H. 2001. Do Germans save to leave an estate? An examination of the bequest motive. *Scandinavian Journal of Economics*, **103**(3), 391–414.

Kassenboehmer, S. C., & Haisken-DeNew, J. P. 2009. You're fired! The causal negative effect of entry unemployment on life satisfaction. *The Economic Journal*, **119**(536), 448–462.

Kazis, L. E., Meenan, R. F., & Anderson, J. J. 1983. Pain in the rheumatic diseases. Investigation of a key health status component. *Arthritis & Rheumatism*, **26**(8), 1017–1022.

Keane, M. 1997. Current issues in discrete choice modeling. *Marketing Letters*, **8**(3), 307–322.

Keane, M., & Wasi, N. 2013. Comparing alternative models of heterogeneity in consumer choice behavior. *Journal of Applied Econometrics*, **28**(6), 1018–1045.

Kirchhoff, T., Ruof, J., Mittendorf, T., Rihl, M., Bernateck, M., Mau, W., Zeidler, H., Schmidt, R. E., & Merkesdal, S. 2011. Cost of illness in rheumatoid arthritis in Germany in 1997-98 and 2002: cost drivers and cost savings. *Rheumatology*, **50**(4), 756–761.

Kirwan, J. R., & Hewlett, S. 2007. Patient perspective: reasons and methods for measuring fatigue in rheumatoid arthritis. *The Journal of Rheumatology*, **34**(5), 1171–1173.

Kløjgaard, M. E., Bech, M., & Søgaard, R. 2011. Designing a stated choice experiment: the value of a qualitative process. *Journal of Choice Modelling*, **5**(2), 1–18.

Kogut, T., & Ritov, I. 2005a. The "identified victim" effect: an identified group or just a single individual? *Journal of Behavioral Decision Making*, **18**(3), 157–167.

Kogut, T., & Ritov, I. 2005b. The singularity effect of identified victims in separate and joint evaluations. *Organizational Behavior and Human Decision Processes*, **97**(2), 106–116.

Krokavcova, M., Nagyoa, I., Rosenberger, J., Gavelova, M., Berrie, M., Gdovinova, Z., Groothoff, J. W., & van Dijk, J. P. 2012. Employment status and perceived health status in younger and older people with multiple sclerosis. *International Journal of Rehabilitation Research*, **35**(1), 40–47.

Kuhfeld, W. F. 2010. *Experimental design: efficiency, coding, and choice designs.* `http://support.sas.com/techsup/technote/mr2010c.pdf`. [Accessed 3 July 2013].

Lamm, R. D. 2001. Compassion of unidentified lives. *Healthplan*, **42**(May/June).

Lancaster, K. J. 1966. A new approach to consumer theory. *The Journal of Political Economy*, **74**(2), 132–157.

Lancsar, E., & Louviere, J. J. 2006. Deleting irrational responses from discrete choice experiments. *Health Economics*, **15**(8), 797–811.

Lancsar, E., Wildman, J., Donaldson, C., Ryan, M., & Baker, R. 2011. Deriving distributional weights for QALYs through discrete choice experiments. *Journal of Health Economics*, **30**(2), 466–478.

Lancsar, E., Louviere, J. J., Donaldson, C., Currie, G., & Burgess, L. 2013. Best worst discrete choice experiments in health: methods and an application. *Social Science & Medicine*, **76**(1), 74–82.

Liu, J. T., Hammitt, J. K., & Liu, J. L. 1997. Estimated hedonic wage function and value of a statistical life in a developing country. *Economics Letters*, **57**(3), 353–358.

Liu, L., & Nelson, W. S. 2006. Endogenous private safety investment and the willingness to pay for mortality risk reductions. *European Economic Review*, **50**(8), 2063–2074.

Llewellyn-Thomas, H., Sutherland, H. J., Tibishirani, R., Ciampi, A., Till, J. E., & Boyd, N. F. 1984. Describing health states: methodologic issues in obtaining values for health states. *Medical Care*, **22**(6), 543–552.

Louviere, J. J. 1973. Theory, methodology and findings in mode choice behaviour. *Working Paper No. 11, The Institute of Urban and Regional Research, The University of Iowa, Iowa City.*

Louviere, J. J. 1988. Conjoint analysis modelling of stated preferences. *Journal of Transport Economics and Policy*, **22**(1), 93–119.

Louviere, J. J., & Hensher, D. A. 1982. On the design and analysis of simulated choice or allocation experiments in travel choice modelling. *Transportation Research Record*, **890**, 11–17.

Louviere, J. J., & Meyer, R. J. 2007. Formal choice models of informal choices: what choice modeling research can (and can't) learn from behavioral theory. *In: Review of Marketing Research.* N. K. Malhotra (ed.), M. E. Sharpe, New York.

Louviere, J. J., & Swait, J. D. 1993. The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research*, **30**(3), 305–314.

Louviere, J. J., & Woodworth, G. 1983. Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. *Journal of Marketing Research*, **20**(4), 350–367.

Louviere, J. J., & Woodworth, G. G. 1990. Best-worst scaling: a model for largest difference judgments. *Working Paper. University of Alberta, Faculty of Business.*

Louviere, J. J., Meyer, R. J., Bunch, D. S., Carson, R., Dellaert, B., Hanemann, W. M., Hensher, D. A., & Irwin, J. 1999. Combining sources of preference data for modelling complex decision processes. *Marketing Letters*, **10**(3), 205–217.

Louviere, J. J., Hensher, D. A., & Swait, J. D. 2000. *Stated choice methods: analysis and applications.* Cambridge University Press: Cambridge.

Louviere, J. J., Street, D. J., Carson, R. T., Ainslie, A., DeShazo, J. R., Cameron, T. A., Hensher, D. A., Kohn, R., & Marley, T. 2002. Dissecting the random component of utility. *Marketing Letters*, **13**(3), 177–193.

Louviere, J. J., Street, D., Burgess, L., Wasi, N., Islam, T., & Marley, A. A. J. 2008. Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information. *Journal of Choice Modelling*, **1**(1), 128–163.

Louviere, J. J., Flynn, T. N., & Carson, R. T. 2010. Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling*, **3**(3), 57–72.

Luce, R. D. 1959. *Individual choice behavior: a theoretical analysis.* Wiley: New York.

Marley, A. A. J., & Louviere, J. J. 2005. Some probabilistic models of best, worst and best-worst choices. *Journal of Mathematical Psychology*, **49**(6), 464–480.

Marley, A. A. J., Flynn, T. N., & Louviere, J. J. 2008. Probabilistic models of set-dependent and attribute-level best-worst choice. *Journal of Mathematical Psychology*, **52**(5), 281–296.

Mas-Colell, A., Whinston, M. D., & Green, J. R. 1995. *Microeconomic theory.* Oxford University Press: New York and Oxford.

May, K. O. 1954. Intransitivity, utility and the aggregation of preference patterns. *Econometrica*, **22**(1), 1–13.

McFadden, D. L. 1974. Conditional logit analysis of qualitative choice behaviour. *In: Frontiers in Economics.* P. Zarembka (ed.). Academic Press: New York.

Meenan, R. F., Gertman, P. M., & Mason, J. H. 1980. Measuring health status in arthritis - the arthritis impact measurement scales. *Arthritis & Rheumatism*, **23**(2), 146–152.

Meenan, R. F., Mason, J. H., Anderson, J. J., Guccione, A. A., & Kazis, L. E. 1992. AIMS2 - The content and properties of a revised and expanded arthritis impact measurement scales health status questionnaire. *Arthritis & Rheumatism*, **35**(1), 1–10.

Merkesdal, S., Ruof, J., Schöfski, O., Bernitt, K., Zeidler, H., & Mau, W. 2001. Indirect medical costs in early rheumatoid arthritis. *Arthritis & Rheumatism*, **44**(3), 528–534.

Miguel, F. S., Ryan, M., & Amaya-Amaya, M. 2005. Irrational stated preferences: a quantitative and qualitative investigation. *Health Economics*, **14**(3), 307–322.

Mill, J. S. 1836. *On the definition of political economy; and on the method of investigation proper to it.* London and Westminster Review.

Minnock, P., Fitzgerald, O., & Bresnihan, B. 2003. Women with established rheumatoid arthritis perceive pain as the predominant impairment of health status. *Rheumatology*, **42**(8), 995–1000.

Moore, R. F. 1996. Caring for identified versus statistical lives: an evolutionary view of medical distributive justice. *Ethology and Sociobiology*, **17**(6), 379–401.

Mrozek, J. R., & Taylor, L. O. 2002. What determines the value of life? A meta-analysis. *Journal of Policy Analysis and Management*, **21**(2), 253–270.

NICE. 2009. *Appraising life-extending, end of life treatments*. `http://www.nice.org.uk/media/E4A/79/SupplementaryAdviceTACEoL.pdf`. [Accessed: 5 July 2012].

Novella, J. L., Jochum, C., Jolly, D., Morrone, I., Ankri, J., Bureau, F., & Blanchard, F. 2001. Agreement between patients' and proxies' reports of quality of life in Alzheimer's disease. *Quality of Life Research*, **10**(5), 443–452.

Olsen, J. A., & Donaldson, C. 1998. Helicopters, hearts and hips: using willingness to pay to set priorities for public sector health care programmes. *Social Science & Medicine*, **46**(1), 1–12.

Peeters, Y., & Stiggelbout, A. M. 2010. Health state valuations of patients and the general public analytically compared: a meta-analytical comparison of patient and population health state utilities. *Value in Health*, **13**(2), 306–309.

Permain, D., Swanson, J., Kroes, E., & Bradley, M. 1991. *Stated preference techniques: a guide to practice*. Steer Davies Gleave and Hague Consulting Group: London.

Persky, J. 1995. The ethology of homo economicus. *The Journal of Economic Perspectives*, **9**(2), 221–231.

Potoglou, D., Burge, P., Flynn, T. N., Netten, A., Malley, J., Forder, J., & Brazier, J. E. 2011. Best-worst scaling vs. discrete choice experiments: an empirical comparison using social care data. *Social Science & Medicine*, **72**(10), 1717–1727.

Pratt, J. W., & Zeckhauser, R. J. 1996. Willingness to pay and the distribution of risk and wealth. *Journal of Political Economy*, **104**(4), 747–763.

Pyne, J. M., Fortney, J. C., Tripathi, S., Feeny, D., Ubel, P., & Brazier, J. 2009. How bad is depression? Preference score estimates from depressed patients and the general population. *Health Services Research*, **44**(4), 1406–1423.

Ratcliffe, J., Flynn, T. N., Sawyer, M., Stevens, K., Brazier, J., & Burgess, L. 2011. Valuing child health utility 9D health states with a young adolescent sample: a feasibility study to compare best-worst scaling discrete-choice experiment, standard gamble and time trade-off methods. *Applied Health Economics and Health Policy*, **9**(1), 15–27.

Revelt, D., & Train, K. 1998. Mixed logit with repeated choices: households' choices of appliance efficiency level. *The Review of Econonomics and Statistics*, **80**(4), 647–657.

Richardson, J., & McKie, J. 2003. The rule of rescue. *Social Science & Medicine*, **56**(12), 2407–2419.

Rosemann, T., Körner, T., Wensing, M., Schneider, A., & Szecsenyi, J. 2005. Evaluation and cultural adaptation of a German version of the AIMS2-SF questionnaire (German AIMS2-SF). *Rheumatology*, **44**(9), 1190–1195.

Ryan, M. 2004. A comparison of stated preference methods for estimating monetary values. *Health Economics*, **13**(3), 291–296.

Ryan, M., & Gerard, K. 2003. Using discrete choice experiments to value health care programmes: current practice and future research reflections. *Applied Health Economics and Health Policy*, **2**(1), 55–64.

Ryan, M., & Watson, V. 2009. Comparing welfare estimates from payment card contingent valuation and discrete choice experiments. *Health Economics*, **18**(4), 389–401.

Salkeld, G., Ryan, M., & Short, L. 2000. The veil of experience: do consumers prefer what they know best? *Health Economics*, **9**(3), 267–270.

Schelling, T. C. 1968. The life you save may be your own. *In: Problems in public expenditure analysis*. The Brookings Institute: Washington DC.

Severin, F., Schmidtke, J., Mühlbacher, A., & Rogowski, W. H. 2013. Eliciting preferences for priority setting in genetic testing: a pilot study comparing best-worst scaling and discrete-choice experiments. *European Journal of Human Genetics*, **21**(11), 1202–1208.

Small, D. A., & Loewenstein, G. 2003. Helping a victim or helping the victim: altruism and identifiability. *The Journal of Risk and Uncertainty*, **26**(1), 5–16.

Small, D. A., Loewenstein, G., & Slovic, P. 2007. Sympathy and callousness: the impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes*, **102**(2), 143–153.

Sossong, Björn. 2012. Rescuing Schelling's girl: revisiting the preference for identified lives using choice analysis. *FOR 655 Working Paper No. 32/2012, Jacobs University Bremen*.

Street, D. J., Burgess, L., & Louviere, J. J. 2005. Quick and easy choice sets: constructing optimal and nearly optimal stated choice experiments. *International Journal of Research in Marketing*, **22**(4), 459–470.

Swain, M. G. 2000. Fatigue in chronic disease. *Clinical Science*, **99**(1), 1–8.

Telser, H., & Zweifel, P. 2007. Validity of discrete-choice experiments evidence for health risk reduction. *Applied Economics*, **39**(1), 69–78.

Theil, H. 1969. A multinomial extension of the linear logit model. *International Economic Review*, **10**(3), 251–259.

Thurstone, L. L. 1927. A law of comparative judgement. *Psychological Review*, **34**(4), 278–286.

Train, K. E. 1986. *Qualitative choice analysis: theory, econometrics, and an application to automobile demand*. MIT Press: Cambridge.

Tsuge, T., Kishimoto, A., & Takeuchi, K. 2005. A choice experiment approach to the valuation of mortality. *Journal of Risk and Uncertainty*, **31**(1), 73–95.

Tversky, A. 1969. Intransitivity of preferences. *Psychological Review*, **76**(1), 31–48.

Ubel, P. A., Richardson, J., & Menzel, P. 2000. Societal value, the person trade-off, and the dilemma of whose values to measure for cost-effectiveness analysis. *Health Economics*, **9**(2), 127–136.

Ubel, P. A., Loewenstein, G., Hershey, J., Baron, J., Mohr, T., Asch, D. A., & Jepson, C. 2001. Do nonpatients underestimate the quality of life associated with chronic health conditions because of a focusing illusion? *Medical Decision Making*, **21**(3), 190–199.

Ubel, P. A., Loewenstein, G., & Jepson, C. 2003. Whose quality of life? A commentary exploring discrepanices between health state evaluations of patiens and the general public. *Quality of Life Research*, **12**(6), 599–607.

Viscusi, W. K., & Aldy, J. E. 2003. The value of a statistical life: a critical review of market estimates throughout the world. *Journal of Risk and Uncertainty*, **27**(1), 5–76.

Whitty, J., Ratcliffe, J., Chen, G., & Scuffham, P. 2013. A comparison of discrete choice and best worst scaling methods to assess Australian public preferences for the funding of new health technologies. *Working Paper presented at the International Choice Modelling Conference 2013.*

Winkelmann, L., & Winkelmann, R. 1998. Why are the unemployed so unhappy? Evidence from panel data. *Economica*, **65**(257), 1–15.

Wolfe, F., Hawley, D. J., & Wilson, K. 1996. The prevalence and meaning of fatigue in rheumatic disease. *The Journal of Rheumatology*, **23**(8), 1407–1417.